Role of OLAP Technology in Data Warehousing for Knowledge Discovery

Sharad Chauhan
Assistant Professor, E-max Group of institutions

Abstract—There are a set of noteworthy newfangled concepts and tools developed into an innovative technology that makes it conceivable to occurrence the problem of providing all the key people in the innovativeness with admittance to whatever level of information needed for the inventiveness to endure and flourish in an progressively modest world. The term that has come to characterize this new technology is “Data Warehousing.” The problem of getting combined and generalized information fast from an active enterprise database becomes actual having its data been accumulated for some years. The classical reports even if optimized for particular purposes do not let one obtain fast the enterprise information with differently data-dependent views. The problem is proper to absolutely all the systems that accumulate large data volumes of information for further processing. To solve the problem is the destiny of the OLAP (On-Line Analytical Processing) technology. The technology nowadays acquiring more and more popularity is assigned to be active and operative handle for multidimensional data and Knowledge Discovery is defined as “the non-trivial extraction of implicit, unknown, and potentially useful information from data.” This paper shows the role of OLAP Technology in Data Warehousing for Knowledge Discovery.

Keywords—Data Warehousing, OLAP Technology, Knowledge Discovery, Multidimensional Data

I. INTRODUCTION

Data Warehousing has grown out of the repeated attempts on the portion of various Researchers and Groups to provide their organizations elastic, effective and efficient means of getting at the sets of data that have come to represent one of the organization’s most serious and valuable properties. Data Warehousing is a field that has grown out of the combination of a number of different technologies and experiences over the last two decades. These involvements have allowed the IT industry to identify the main problems that have to be explained.

In simple words, a Data Warehouse refers to a database that is sustained separately from an organization’s operative databases. According to William H. Inmon, a leading architect in the construction of data warehouse systems, “A Data Warehouse is a subject-oriented, incorporated, time-variant, and non-volatile collection of data in support of management’s decision making process.” Here, by this definition, the four major keywords Subject-oriented, Integrated, Time-variant and Non-volatile, distinguish data warehouses from other data repository systems, such as RDBMS, TPS, and file systems.

According to the above believed theory, we get Data Warehousing is the process of building and using data warehouses. The production of a data warehouse requires data cleaning, data integration and data consolidation.

1.1.1 How are Organizations using the information from Data Warehouses?

Many organizations use the data to support business decision-making activities, including:

- Increasing Client attention
- Relocation – products and managing product groups
- Examining Operations and looking for sources of profit
- Managing the Customer relationships
- Making Environmental Corrections

1.1.2 Three-Tier data Warehouse Architecture

Data Warehouses often adopt a three-tier architecture, as shown in figure given below:

![Figure 1](image_url)
Permitting to the above figure, there are three level of Data Warehouse Architecture. These are (from Low to High): Bottom Tier, also known as Data Warehouse Server, Middle Tier, also known as OLAP Server, and Top Tier, also known as Front End Tools.

1.1.3 Data Warehouse Models:
From the Architecture point of sight, there are three Data Warehouse Models: the Enterprise Warehouse, the Data Mart, and the Virtual Warehouse.

II. ONLINE ANALYTICAL PROCESSING (OLAP) TECHNOLOGY

Online Analytical Processing (OLAP) has appeared as a “advance” technology that can provide the base for EIS solutions. Using OLAP, senior managers are able to view hundreds of realistic and tabular displays that present a conception of their institution’s business process. Because OLAP technology provides user and Data Scalability, Performance, Read/Write Capabilities and Calculation Functionality, it meets all the requirements of a data mart. Two other options — Personal Productivity Tools, and Data Query and Reporting Tools—cannot provide the same level of support. Personal productivity tools such as spreadsheets and statistical packages reside on individual PCs, and therefore support only small amounts of data to a single user.

2.1 Features of OLAP Technology.

Most important in OLAP technology are its sophisticated analytic capabilities. Main features of OLAP Technology including:

- Aggregations, which simply add numbers based upon levels defined by the application. For example, the application may call for adding up sales by week, month, quarter and year.
- Matrix Calculations, which are similar to calculations executed within a standard spreadsheet. For example, variances and ratios are matrix calculations.
- Cross-Dimensional Calculations, which are similar to the calculations executed when spreadsheets are linked and formulas combine cells from different sheets. A percent product share calculation is a good example of this, as it requires the summation of a total and the calculation of percentage contribution to total sales of a given product.

2.2 Categories of OLAP Technology.

OLAP technology may be either Relational or Multidimensional in nature.

Relational OLAP Technologies, basically suitable for large and detail-level sets of data. But, these technologies have inherent weaknesses in a decision- support environment. Response time for decision-support queries in a relational framework can vary from minutes to hours. Calculations are limited to aggregations and simple matrix processing. Changes to metadata structures—for example, the organization of sales territories—usually require manual administrator intervention and re-creation of all summary tables.

On the other hand, Multidimensional Technology is free from the limitations that relational databases face in decision-support environments, as multidimensional OLAP delivers sub-second response times while supporting hundreds and thousands of concurrent users.

2.3 How OLAP Technology works.

How does it work? OLAP technology contains of two major components, the Server and the Client. Naturally Server is a Multi-User, LAN based database that is overloaded either from legacy systems or from Data Warehouse. OLAP’s visualization will tell designs of business process that are hidden in the data. The server Think of OLAP databases as multi-dimensional arrays or dices of data—capable of holding hundreds of thousands of rows and columns of both text and numbers. The current terminology for these record servers is multi-dimensional databases (MDDs).

III. KNOWLEDGE DISCOVERY

Knowledge Discovery is defined as “the non-trivial removal of implicit, hidden, and potentially valuable information from data”. There are many Knowledge Discovery Methods in use and under development. Some of these techniques are general, while others are domain-specific.

3.1 Basic features to various Knowledge Discovery Techniques:

The following are basic features that are common to various Knowledge Discovery Techniques:

- All methods deal with large volumes of data. Large volumes of data are required to provide sufficient information to derive additional knowledge
- Effectiveness is required due to volume of data
- Correctness is required to assure that discovered information is valid
- Use of a high-level language i.e. the results should be obtainable in a manner that is clear to individuals
- All methods use some form of automatic knowledge
- All produce some remarkable results

Knowledge Discovery provides the ability to discover new and meaningful data by using existing data. Knowledge Discovery rapidly exceeds the human capacity to analyze large data sets.

3.2 Knowledge Discovery Techniques.

There are many different approaches that are classified as Knowledge Discovery Techniques. These are:

- Measurable Approaches, such as the Probabilistic and Statistical Approaches
- Classification Methods such as Bayesian Classification, Inductive Logic, Data Cleaning / Pattern Finding, and Decision Tree Analysis.
- Additional Approaches include Deviation and Trend Analysis, Genetic Algorithms, Neural Networks, and Hybrid Approaches that combine two or more techniques.

Due to these techniques can be used and combined, there is a absence of contract on how these techniques should be categorized. For example, the Bayesian Approach may be logically grouped with Probabilistic Approaches.

3.2.1 Quantitative Approaches.

3.2.1.1 Probabilistic Approach.
This class of Knowledge Discovery Techniques uses Graphical representation models to compare different data representations. These models are built on probabilities and data independencies. These approaches are useful for applications involving uncertainty and applications organized such that a probability may be assigned to each “outcome” or bit of discovered knowledge. Probabilistic techniques may be used in analytical structures and in planning and control systems. Automated probabilistic tools are available both commercially and in the public domain.

3.2.1.2 Statistical Approach
The Statistical Approach uses rule discovery and is based on data relationships. This type of induction is used to generalize patterns in the data and to build rules from the well-known patterns. Online analytical processing (OLAP) is an example of a Statistically-Oriented Approach. Automatic statistical tools are open for commercially and in the user domain.

3.2.2 Visualization Utilization Methods
The Visualization Utilization Methods utilizes Visualization techniques of Knowledge Discovery.

3.2.3 Classification Method
Classification is probably the oldest and most widely-used of all the Knowledge Discovery Approaches. This approach collects the data according to similarities or sessions. There are many types of classification techniques and numerous automated tools available.

3.2.3.1 Bayesian Approach to Knowledge Discovery “is a graphical model that uses an directed Acyclic Graph”. Although the Bayesian method uses probabilities and a graphical means of representation, it is also measured as a type of classification. Bayesian networks are typically used when the ambiguity is associated with an outcome that can be stated in terms of a probability. This approach based on encoded domain knowledge and has been used for analytical systems.

3.2.3.2 Pattern Discovery and Data Cleaning is another type of classification that analytically reduces a large database to a few relevant and useful records. If redundant and uninteresting data is eliminated, the task of determining patterns in the data is simplified. The pattern discovery and data cleaning techniques are useful for reducing huge volumes of application data, such as those encountered when examining automatic sensor recordings.

3.2.4 Another Approaches

3.2.4.1 Deviation and Trend Analysis: Pattern detection by clarifying important trends is the basis for this Knowledge Discovery Method. Deviation and Trend analysis methods are normally applied to temporal records. A good application for this type of Knowledge Discovery is the analysis of traffic on large communications networks.

3.2.4.2 Neural Networks may be used as a method of knowledge discovery. Neural networks are particularly valuable for pattern recognition, and are sometimes grouped with the classification approaches.

3.2.4.3 Genetic Algorithms also used for classification, are similar to neural networks although they are typically considered more powerful. There are tools for the genetic approach available commercially.

IV. CONCLUSIONS
Knowledge discovery provides the capability to discover new and meaningful information by using existing data. Knowledge discovery quickly exceeds the human capacity to analyze large data sets. OLAP Technology, basically, is the solution of the problem getting combined and generalized information fast from an active enterprise database i.e. from data warehouses. With a wide plethora of applications across different Industries, OLAP Technology is poised to become a field that will attract significant investment and research in the coming years. Most of the Organizations in today’s world have invested vast amounts of capital and resources and adopted technology to collect and store huge amounts information about their business. However, implementing OLAP Technology for Knowledge Discovery is also not that easy. Groups wanting to implement OLAP Technology for Knowledge Discovery have to grapple with issues such as approach, knowledge, organizational culture and information of complete Data Warehouses. And, needless to say, OLAP Technology will have a key role to play in Data Warehousing for Knowledge Discovery.

REFERENCES

[1]. Jiawei Han (University of Illinois at Urbana-Champaign) & Micheline Kamber. Data Mining : Concepts and Techniques 2/e, Morgan Kaufmann Publishers – An Imprint of Elsevier, 500 Sansome Street, Suite 400, San Francisco, CA, 2006, pp. 105-150

[2]. Data Warehousing Technology- A White Paper by Ken Orr, The Ken Orr Institute, 2921 S.W. Wanamaker Drive, Topeka, KS 66614,

[3]. The Data Warehousing Information Centre, Larry Greenfield, LGI Systems Incorporated.
