ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

OLAP Technology in Data Warehouses

Dr. Atul Khurana

Department of Computer Science & Engineering, Aryabhatta Group

Abstract: Recently, a set of significant new concepts and tools have evolved into a new technology that makes it possible to attack the problem of providing all the key people in the enterprise with access to whatever level of information needed for the enterprise to survive and prosper in an increasingly competitive 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.

1. DATA WAREHOUSE – INTRODUCTION

1.1 What is Data Warehouse?

Data Warehousing has grown out of the repeated attempts on the part of various Researchers and Organizations to provide their organizations flexible, effective and efficient means of getting at the sets of data that have come to represent one of the organization's most critical and valuable assets. Data Warehousing is a field that has grown out of the integration of a number of different technologies and experiences over the last two decades. These experiences have allowed the IT industry to identify the key problems that have to be solved. Data Warehousing provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions. Data warehouse systems are valuable tools in today's competitive, fast-evolving world. Data Warehouses generalize and consolidate data in multidimensional space. The construction of data warehouses involves data cleaning, data integration, and data transformation.

In simple words, a Data Warehouse refers to a database that is maintained separately from an organization's operational databases [Jiawei Han & Micheline Kamber. 2006].

According to William 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" [Jiawei Han & Micheline Kamber. 2006]. 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 said theory, we get Data Warehousing is the process of constructing and using data warehouses. The construction of a data warehouse requires data cleaning, data integration and data consolidation. The utilization of a data warehouse often necessitates a collection of decision support technologies.

1.1.1 How are Organizations using the information from Data Warehouses?

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

- Increasing Customer focus
- Repositioning –products and managing product portfolios
- Analyzing Operations and looking for sources of profit
- Managing the Customer relationships
- Making Environmental Corrections
- Managing the Cost of corporate assets

1.1.2 Three-Tier data Warehouse Architecture

Data Warehouses often adopt a three-tier architecture [Jiawei Han & Micheline Kamber. 2006], as shown in figure given below:

According to the 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.

The Bottom Tier is a Warehouse Database Server that is almost always a relational database system. Back-end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources. These tools & utilities perform data extraction, cleaning and transformation, as well as load and refresh functions to update the data warehouse. The data are extracted using application program interfaces known as gateways (e.g., ODBC – Open Database Connection, OLEDB – Open Linking and Embedding for Databases, JDBC – Java

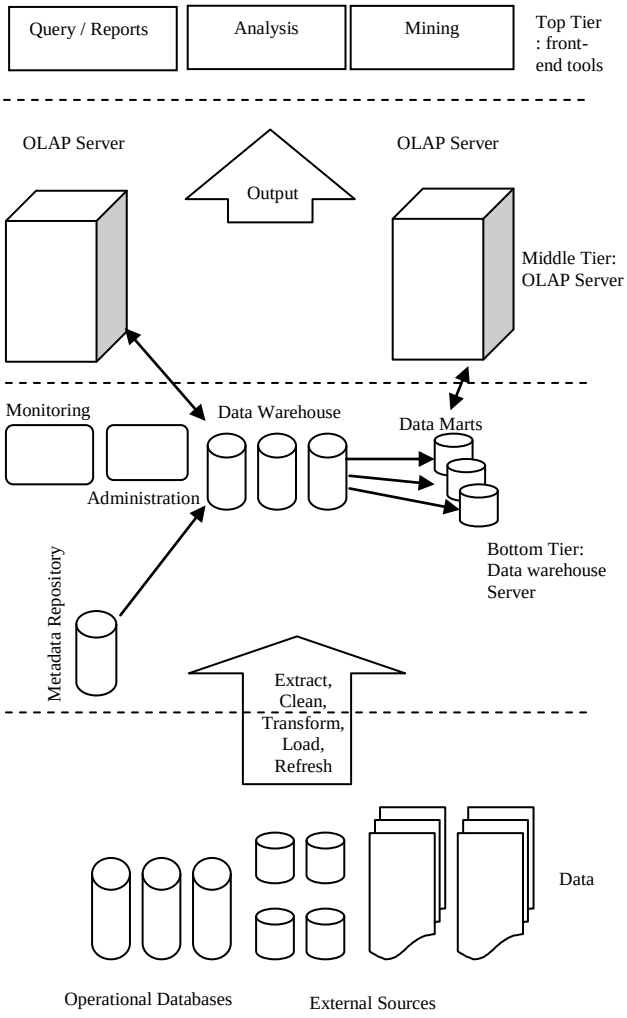
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

Database Connection). This tier also contains a metadata repository, which stores information about the data warehouse and its contents.

The Middle Tier is an OLAP Server that is typically implemented using either a) 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.

The Top Tier is a front-end client layer, which contains query and reporting tools, analysis tools, and/or data mining tools (e.g., Trend Analysis, Prediction, and so on).



1.1.3 Data Warehouse Models:

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

2. ONLINE ANALYTICAL PROCESSING (OLAP) TECHNOLOGY.

Online Analytical Processing (OLAP) has emerged as a "breakthrough" technology that can provide the foundation for EIS solutions. Using OLAP, senior managers are able to view hundreds of graphic and tabular displays that present a visualization 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. Data query and reporting tools are SQL-driven, and frequently used for list-oriented, basic drill-down analysis and report generation. These tools do not offer the predictable performance or robust calculations of OLAP. The OLAP technology option supports collaboration throughout the business management cycle of reporting, analysis, what-if modeling and planning.

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.

Procedural Calculations, in which specific calculation rules are defined and executed in a specific order. For example, allocating advertising expense as a percent of revenue contribution per product is a procedural calculation, requiring procedural logic to properly model and execute sophisticated business rules that accurately reflect the business.

OLAP-Aware Calculations, which provide the analytical intelligence necessary for multi-dimensional analysis, such as the understanding of hierarchy relationships within dimensions. These calculations include time intelligence and financial intelligence. For example, an OLAP-aware calculation would calculate inventory balances in which Q1 ending inventory is understood not to be the sum of January, February and March inventories.

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

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

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. In addition, it supports the full range of calculations, from aggregations to procedural calculations.

Data warehousing has traditionally focused on relational technology. While well-suited to managing transactions and storing large amounts of data, relational databases are typically unable to handle ad hoc, speed-of-thought analytical querying for large user communities. Online analytical processing (OLAP) technology, however, provides the scalability, performance and analytic capabilities necessary to support sophisticated, calculation-intensive queries for large user populations. For these reasons, relational and OLAP technologies are often combined for maximum benefits.

2.3 How OLAP Technology works.

How does it work? OLAP technology consists of two major components, the Server and the Client. Typically the Server is a Multi-User, LAN based database that is loaded either from legacy systems or from Data Warehouse. We don't need a Data Warehouse in order to implement OLAP, but if we have historical data, OLAP's visualization will reveal patterns of business process that are hidden in the data.

The server Think of OLAP databases as multi-dimensional arrays or cubes of data—actually cubes of cubes—capable of holding hundreds of thousands of rows and columns of both text and numbers. The current terminology for these database servers is multi-dimensional databases (MDDs). The MDDs are loaded from data source (legacy or warehouse) according to an aggregation model that we define.

Fortunately, defining the model and loading the database can be very easy; for some OLAP products, no programming is required to build the model or to load the data.The client component for several OLAP products presents a spreadsheet-type interface.

3. KNOWLEDGE DISCOVERY

Knowledge Discovery is defined as "the non-trivial extraction of implicit, unknown, and potentially useful information from data". It is a growing field.

There are many Knowledge Discovery Methodologies in use and under development. Some of these techniques are generic, while others are domain-specific.

3.1 Common Basic features to various Knowledge Discovery Techniques:

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

- All approaches deal with large amounts of data. Large amounts of data are required to provide sufficient information to derive additional knowledge
- Efficiency is required due to volume of data
- Accuracy is required to assure that discovered knowledge is valid
- Use of a high-level language i.e. the results should be presented in a manner that is understandable by humans
- All approaches use some form of automated learning
- All produce some interesting results

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. The amount of data that requires processing and analysis in a large database exceeds human capabilities, and the difficulty of accurately transforming raw data into knowledge surpasses the limits of traditional databases. Therefore, the full utilization of stored data depends on the use of knowledge discovery techniques.

The usefulness of future applications of Knowledge Discovery is far-reaching. It may be used as a means of information retrieval, in the same manner that intelligent agents perform information retrieval on the web. New patterns or trends in data may be discovered using these techniques. Knowledge Discovery may also be used as a basis for the intelligent interfaces of tomorrow, by adding a knowledge discovery component to a database engine or by integrating Knowledge Discovery with spreadsheets and visualizations.

3.2 Knowledge Discovery Techniques.

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

- Quantitative Approaches, such as the Probabilistic and Statistical Approaches
- Visualization Utilization Approaches
- Classification Approaches such as Bayesian Classification, Inductive Logic, Data Cleaning / Pattern Discovery, and Decision Tree Analysis.
- Other Approaches include Deviation and Trend Analysis, Genetic Algorithms, Neural Networks, and Hybrid Approaches that combine two or more techniques.

Because of the ways that these techniques can be used and combined, there is a lack of agreement on how these techniques should be categorized.
For example, the Bayesian Approach may be logically grouped with Probabilistic Approaches, Classification Approaches, or Visualization Approaches. For the sake of organization, each approach described here is included in the group that it seemed to fit best. However, this selection is not intended to imply a strict categorization.

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

3.2.1 Quantitative Approaches.

3.2.1.1 Probabilistic Approach.

This family of Knowledge Discovery Techniques utilizes Graphical representation models to compare different knowledge representations. These models are based on probabilities and data independencies. They are useful for applications involving uncertainty and applications structured such that a probability may be assigned to each "outcome" or bit of discovered knowledge. Probabilistic techniques may be used in diagnostic systems 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. An "inductive learning algorithm can automatically select useful join paths and attributes to construct rules from a database with many relations". This type of induction is used to generalize patterns in the data and to construct rules from the noted patterns. Online analytical processing (OLAP) is an example of a Statistically-Oriented Approach. Automated statistical tools are available both commercially and in the public domain.

3.2.2 Visualization Utilization Approaches

The Visualization Utilization Approaches utilizes Visualization techniques of Knowledge Discovery.

3.2.3 Classification Approach

Classification is probably the oldest and most widely-used of all the Knowledge Discovery Approaches. This approach groups data according to similarities or classes. 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 directed arcs exclusively to form an directed Acyclic Graph". Although the Bayesian approach uses probabilities and a graphical means of representation, it is also considered a type of classification. Bayesian networks are typically used when the uncertainty associated with an outcome can be expressed in terms of a probability. This approach relies on encoded domain knowledge and has been used for diagnostic systems. Other pattern recognition applications, including the Hidden Markov Model, can be modeled using a Bayesian approach. Automated tools are available both commercially and in the public domain.

3.2.3.2 Pattern Discovery and Data Cleaning is another type of classification that systematically reduces a large database to a few pertinent and informative records. If redundant and uninteresting data is eliminated, the task of discovering patterns in the data is simplified. This approach works on the premise of the old adage, "less is more". The pattern discovery and data cleaning techniques are useful for reducing enormous volumes of application data, such as those encountered when analyzing automated sensor recordings. Once the sensor readings are reduced to a manageable size using a data cleaning technique, the patterns in the data may be more easily recognized. Automated tools using these techniques are available both commercially and in the public domain.

3.2.3.3 Decision Tree Approach uses production rules, builds a directed acyclical graph based on data premises, and classifies data according to its attributes. This method requires that data classes are discrete and predefined.

According to, the primary use of this approach is for predictive models that may be appropriate for either classification or regression techniques. Tools for decision tree analysis are available commercially and in the public domain.

3.2.4 Other Approaches

3.2.4.1 Deviation and Trend Analysis: Pattern detection by filtering important trends is the basis for this Knowledge Discovery Approach. Deviation and Trend analysis techniques are normally applied to temporal databases. A good application for this type of Knowledge Discovery is the analysis of traffic on large telecommunications networks.

AT&T uses such a system to locate and identify circuits that exhibit deviation (faulty behavior). The sheer volume of data requiring analysis makes an automated technique imperative. Trend-type analysis might also prove useful for astronomical and oceanographic data, as they are time-based and voluminous. Public domain tools are available for this approach.

3.2.4.2 Neural Networks may be used as a method of knowledge discovery. Neural networks are particularly useful for pattern recognition, and are sometimes grouped with the classification approaches. There are tools available in the public domain and commercially.

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.

3.2.4.4 Hybrid Approach to Knowledge Discovery combines more than one approach and is also called a multi-paradigmatic approach. Although implementation may be more difficult, hybrid tools are able to combine the strengths of various approaches. Some of the commonly used methods combine visualization techniques, induction, neural networks, and rule-based systems to achieve the desired knowledge discovery. Deductive databases and genetic algorithms have also been used in hybrid approaches. There are hybrid tools available commercially and in the public domain.

DOI 10.17148/IJARCCE.2015.4864

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

4. CONCLUSION

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. OLAP Technology as a concept is very attractive and too many organizations is trendy and nice to be associated with. However, implementing OLAP Technology for Knowledge Discovery is also not that easy. Organizations wanting to implement OLAP Technology for Knowledge Discovery have to grabble with issues such as strategy, technology, organizational culture and knowledge 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.

[4] Brachman, R.J., and Anand, T. The Process Of Knowledge Discovery In Databases: A Human-Centered Approach. In Advances In Knowledge Discovery And Data Mining , eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI Press/The MIT Press, Menlo Park, CA., 1996, pp. 37-57.

[5] Buntine, W. Graphical Models For Discovering Knowledge. In Advances In Knowledge Discovery And Data Mining, eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI Press/The MIT Press, Menlo Park, CA., 1996, pp. 59-82.

[6] Frawley, W.J., Piatetsky-Shapiro, G., and Matheus, C. Knowledge Discovery In Databases: An Overview. In Knowledge Discovery In Databases, eds. G. Piatetsky-Shapiro, and W. J. Frawley, AAAI Press/MIT Press, Cambridge, MA., 1991, pp. 1-30.