Data Warehousing, Data Mining, OLAP and OLTP Technologies Are Indispensable Elements to Support Decision-Making Process in Industrial World

Amandeep Kour

Assistant Professor, "Department of Computer Science and Engineering"
M.B.S College of Engineering & Technology, Babliana, Jammu (J&K) India

Abstract- This paper provides an overview of Data warehousing, Data Mining, OLAP, OLTP technologies, exploring the features, new applications and the architecture of Data Warehousing and 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. 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. An OLAP system is market-oriented and is used for data analysis by knowledge workers, including managers, executives and analysts. Data warehousing and OLAP have emerged as leading technologies that facilitate data storage, organization and then, significant retrieval. Decision support places some rather different requirements on database technology compared to traditional on-line transaction processing applications.

Index Terms- Data Warehousing, OLAP, OLTP, Data Mining, Decision Making and Decision Support, Data mining, Data marts, Meta data, ETL (Extraction, Transportation, transformation and loading), Server, Data warehouse architecture.

I. Introduction

Different people have different definitions for a data warehouse. The most popular definition came from "Bill Inmon", who provided the following:

A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process.

Subject-Oriented: A data warehouse can be used to analyze a particular subject area. For example, "sales" can be a particular subject.

Integrated: A data warehouse integrates data from multiple data sources. For example, source A and source B may have different ways of identifying a product, but in a data warehouse, there will be only a single way of identifying a product.

Time-Variant: Historical data is kept in a data warehouse. For example, one can retrieve data from 3 months, 6 months, 12 months, or even older data from a data warehouse. This contrasts with a transactions system, where often only the most recent data is kept. For example, a transaction system may hold the most recent address of a customer, where a data warehouse can hold all addresses associated with a customer.

Non-volatile: Once data is in the data warehouse, it will not change. So, historical data in a data warehouse should never be altered.

Ralph Kimball provided a more concise definition of a data warehouse: A data warehouse is a copy of transaction data specifically structured for query and analysis. This is a functional view of a data warehouse. Kimball did not address how the data warehouse is built like Inmon did; rather he focused on the functionality of a data warehouse.

Data warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker (executive, manager, 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).

II. Data Warehousing

2.1 Definition of data warehousing:

A single, complete and consistent store of data obtained from a variety of different sources made available to end users in what they can understand and use in a business context.

Data Warehousing is defined in many different ways, but not rigorously.

A decision support database that is maintained separately from the organization's operational database Support information processing by providing a solid platform of consolidated, historical data for analysis. Data warehousing is the process of constructing and using data warehouses.Organized around major subjects, such as customer, product, sales Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.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.

2.2 Explains how to design and manage data warehouse systems focusing on project management aspects:

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 OLAP Council (http://www.olapcouncil.org) is a good source of information on standardization efforts across the industry. The symbolic dimension abandons the assumptions of rational behavior and views organizations as some kind of theatres.. Finally, a good source of references on data warehousing and OLAP is the Data Warehousing Information Center.

2.3 Data warehouse Architecture

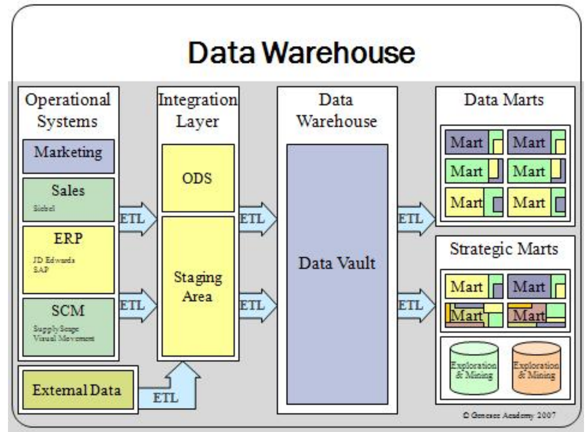

Fig: 1 shows a Data warehousing architecture

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 warehouseservers, 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.

Enterprise warehouse collects all of the information about subjects spanning the entire organization

Data Mart is a subset of corporate-wide data that is of value to a specific groups of users. Its scope is confined to specific, selected groups, such as marketing data mart

Independent vs. dependent (directly from warehouse) data mart.

Virtual warehouse is a set of views over operational databases

Only some of the possible summary views may be materialized.

Meta data is the data defining warehouse objects. It stores Description of the structure of the data warehouse schema, view, dimensions, hierarchies, derived data defined, data mart locations and contents

Operational meta-data is a data lineage (history of migrated data and transformation path), currency of data (active, archived, or purged), monitoring information (warehouse usage statistics, error reports, audit trails)

The algorithms used for summarization and the mapping from operational environment to the data warehouse data related to system performance Warehouse schema, view and derived data definitions.

III. 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. 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 automatic discovery of relationships in typically large database and, in some instances, the use of the discovery results in predicting relationships.
- An essential process where intelligent methods are applied in order to extract data patterns.
- Data mining lets you be proactive
  - Prospective rather than Retrospective

1.1 Why mine data?

Commercial viewpoint…

- Lots of data is being collected and warehoused.
- Computing has become affordable.
- Competitive Pressure is Strong
  - Provide better, customized services for an edge.
  - Information is becoming product in its own right.

1.2 Why Mine Data?

Scientific Viewpoint…

- Data collected and stored at enormous speeds
  - Remote sensor on a satellite
  - Telescope scanning the skies
  - Microarrays generating gene expression data
  - Scientific simulations generating terabytes of data
- Traditional techniques are infeasible for raw data
- Data mining for data reduction
  - Cataloging, classifying, segmenting data
  - Helps scientists in Hypothesis Formation.

4.3 Major Data Mining Tasks

- Classification: Predicting an item class.
- Association Rule Discovery: descriptive.
- Clustering: descriptive, finding groups of items.
- Sequential Pattern Discovery: descriptive.
- Deviation Detection: predictive, finding changes.
- Forecasting: predicting a parameter value
- Description: describing a group.
- Link analysis: finding relationships and associations.

4.3.1 Classification: Definition

- Given a collection of records(training set)
  - Each record contains a set of attributes, one of the attributes is the class.
- Find a model for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A test set is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

4.3.1.1 Classification: Application

- Direct Marketing
  - Goal: Reduce cost of mailing by targeting a set of customers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This {buy, don't buy} decision forms the class attribute.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

4.3.1.2 Associations

- $I = \{i_1, i_2, \ldots i_m\}$: a set of literals, called items.
- Transaction d: a set of items such that $d \subseteq I$
- Database D: a set of transactions
- A transaction d contains X, a set of some items in L, if X Íd.
- An association rule is an implication of the form $X \Rightarrow Y$, where X, $Y \subset I$.

4.3.1.3 Association rules

- Used to find all rules in a basket data
- Basket data also called transaction data
- analyze how items purchased by customers in a shop are related
- discover all rules that have:-
  - support greater than min sup specified by user
  - confidence greater than min conf specified by user
- Example of transaction data:-
  - CD player, music's CD, music's book
  - CD player, music's CD
  - music's CD, music's book
  - CD player
- Let $I = \{i_1, i_2, \ldots i_m\}$ be a total set of items
  D a set of transactions
  d is one transaction consists of a set of items
  - $d \subseteq \mathbf{I}$
- Association rule:-
  - Let $I = \{i_1, i_2, \ldots i_m\}$ be a total set of items
  - D a set of transactions
  - d is one transaction consists of a set of items
  - $d \subseteq \mathbf{I}$
  -

Association rule:-

- $X \rightarrow Y$ where $X \subset I$,$Y \subset I$ and $X \cap Y = \varnothing$
- support = (#of transactions contain $X \cup Y$) / D
- confidence = (#of transactions contain $X \cup Y$) /
- #of transactions contain X .

4.3.1.4 Clustering

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.

4.3.1.4 Clustering Applications

- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
- Approach:
  - Collect different attributes of customers based on their geographical and lifestyle related information
  - Find clusters of similar customers.
  - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

V. Bayesian Belief Networks

Bayesian belief network (also known as Bayesian network, probabilistic network): allows class conditional independencies between subsets of variables

- Two components: (1) A directed acyclic graph (called a structure) and (2) a set of conditional probability tables (CPTs).
- A (directed acyclic) graphical model of causal influence relationships.
- Represents dependency among the variables.
- Gives a specification of joint probability distribution.

1.3 How Are Bayesian Networks Constructed?

- **Subjective construction**: Identification of (direct) causal structure.
- People are quite good at identifying direct causes from a given set of variables & whether the set contains all relevant direct causes.
- Markovian assumption: Each variable becomes independent of its non-effects once its direct causes are known
- E.g., S ‹— F —› A ‹— T, path S—›A is blocked once we know F—›A
- HMM (Hidden Markov Model): often used to model dynamic systems whose states are not observable, yet their outputs are:
- **Synthesis from other specifications**

E.g., from a formal system design: block diagrams & info flow

- **Learning from data**

E.g., from medical records or student admission record.

- Learn parameters give its structure or learn both structure and parms
- Maximum likelihood principle: favors Bayesian networks that maximize the probability of observing the given data set.

5.2 Training Bayesian Networks: Several Scenarios.

- Scenario 1: Given both the network structure and all variables observable: *compute only the CPT entries*
- Scenario 2: Network structure known, some variables hidden: *gradient descent* (greedy hill-climbing) method, i.e., search for a solution along the steepest descent of a criterion function
- Weights are initialized to random probability values
- At each iteration, it moves towards what appears to be the best solution at the moment, w.o. backtracking
- Weights are updated at each iteration & converge to local optimum
- Scenario 3: Network structure unknown, all variables observable: search through the model space to *reconstruct network topology* .
- Scenario 4: Unknown structure, all hidden variables: No good algorithms known for this purpose
- D. Heckerman. [A Tutorial on Learning with Bayesian Networks](). In Learning in Graphical Models, M. Jordan, ed. MIT Press, 1999.

1.4 Neuron: A Hidden/Output Layer Unit

Fig: 2 Hidden/output layer diagram

For Example

$$y = \text{sign}\left(\sum_{i=0}^{n} w_i x_i - \mu_k\right)$$

- An *n*-dimensional input vector **x** is mapped into variable y by means of the scalar product and a nonlinear function mapping

The inputs to unit are outputs from the previous layer. They are multiplied by their corresponding weights to form a weighted sum, which is added to the bias associated with unit. Then a nonlinear activation function is applied to it.

1.5 Genetic Algorithms (GA)

- Genetic Algorithm: based on an analogy to biological evolution
- An initial population is created consisting of randomly generated rules
- Each rule is represented by a string of bits
- E.g., if A_1 and $\neg A_2$ then C_2 can be encoded as 100
- If an attribute has k > 2 values, k bits can be used
- Based on the notion of survival of the fittest, a new population is formed to consist of the fittest rules and their offspring
- The fitness of a rule is represented by its classification accuracy on a set of training examples
- Offspring are generated by crossover and mutation
- The process continues until a population P evolves when each rule in P satisfies a pre-specified threshold
- Slow but easily parallelizable.

5.4.1 Rough Set Approach:

- Rough sets are used to approximately or "roughly" define equivalent classes
- A rough set for a given class C is approximated by two sets: a lower approximation (certain to be in C) and an upper approximation (cannot be described as not belonging to C)
- Finding the minimal subsets (**reducts**) of attributes for feature reduction is NP-hard but a **discernibility matrix** (which stores the differences between attribute values for each pair of data tuples) is used to reduce the computation intensity.

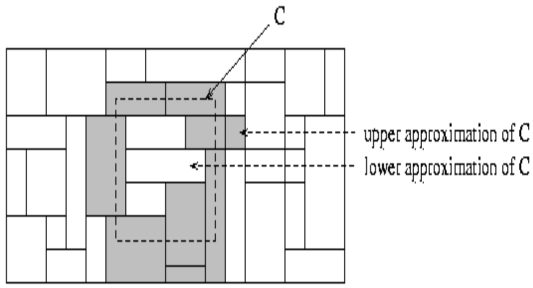

Fig: 3 rough set Approach

VI. Active Learning

- Class labels are expensive to obtain
- Active learner: query human (oracle) for labels
- Pool-based approach: Uses a pool of unlabeled data
  - L: a small subset of D is labeled, U: a pool of unlabeled data in D
  - Use a query function to carefully select one or more tuples from U and request labels from an oracle (a human annotator)
  - The newly labeled samples are added to L, and learn a model
  - Goal: Achieve high accuracy using as few labeled data as possible
- Evaluated using learning curves: Accuracy as a function of the number of instances queried (# of tuples to be queried should be small)
- Research issue: How to choose the data tuples to be queried?

ISSN 2250-3153

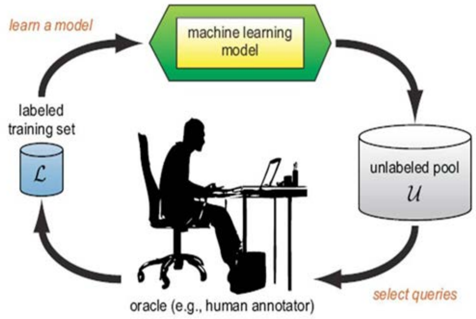

Fig: 7 Rough set Approach

- Uncertainty sampling: choose the least certain ones
- Reduce *version space*, the subset of hypotheses consistent w. the training data
- Reduce expected entropy over U: Find the greatest reduction in the total number of incorrect predictions.

VII. Transfer Learning: Conceptual Framework

- Transfer learning: Extract knowledge from one or more source tasks and apply the knowledge to a target task
- Traditional learning: Build a new classifier for each new task
- Transfer learning: Build new classifier by applying existing knowledge learned from source tasks.

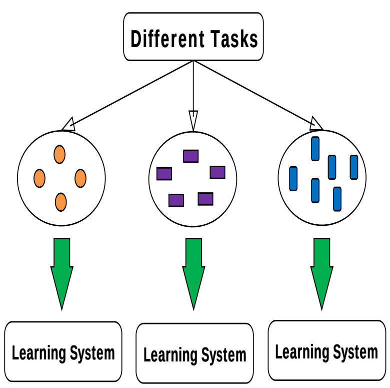

Fig: 8 Traditional Learning Framework

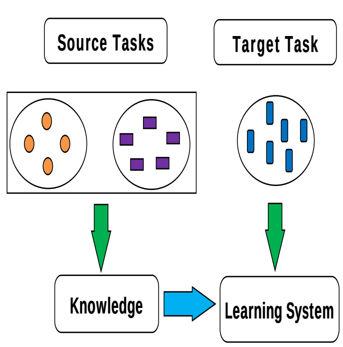

Fig: 9 Transfer Learning Framework

VIII. A Closer Look at CMAR

- **CMAR** (Classification based on Multiple Association Rules: Li, Han, Pei, ICDM'01)
- Efficiency: Uses an enhanced FP-tree that maintains the distribution of class labels among tuples satisfying each frequent itemset
- Rule pruning whenever a rule is inserted into the tree
  - Given two rules, R_1 and R_2, if the antecedent of R_1 is more general than that of R_2 and $conf(R_1) \geq conf(R_2)$, then prune R_2
  - Prunes rules for which the rule antecedent and class are not positively correlated, based on a χ^2 test of statistical significance
- Classification based on generated/pruned rules
  - If only *one rule* satisfies tuple X, assign the class label of the rule
  - If a *rule set* S satisfies X, CMAR
    - divides S into groups according to class labels
    - uses a weighted χ^2 measure to find the strongest group of rules, based on the statistical correlation of rules within a group
    - assigns X the class label of the strongest group.

References

[1] C. M. Bishop, Neural Networks for Pattern Recognition. Oxford University Press, 1995
[2] C. J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2(2): 121-168, 1998
[3] H. Cheng, X. Yan, J. Han, and C.-W. Hsu, Discriminative Frequent pattern Analysis for Effective Classification, ICDE'07
[4] H. Cheng, X. Yan, J. Han, and P. S. Yu, Direct Discriminative Pattern Mining for Effective Classification, ICDE'08
[5] N. Cristianini and J. Shawe-Taylor, Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, 2000
[6] A. J. Dobson. An Introduction to Generalized Linear Models. Chapman & Hall, 1990
[7] G. Dong and J. Li. Efficient mining of emerging patterns: Discovering trends and differences. KDD'99
[8] R. O. Duda, P. E. Hart, and D. G. Stork. Pattern Classification, 2ed. John Wiley, 2001
[9] T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, 2001
[10] S. Haykin, Neural Networks and Learning Machines, Prentice Hall, 2008
[11] D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning, 1995.
[12] V. Kecman, Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic, MIT Press, 2001
[13] W. Li, J. Han, and J. Pei, CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules, ICDM'01
[14] T.-S. Lim, W.-Y. Loh, and Y.-S. Shih. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Machine Learning, 2000

AUTHORS

First Author – AMANDEEP KOUR, Assistant Professor, "Department of Computer Science and Engineering", M.B.S College of Engineering & Technology, Babliana, Jammu (J&K) India, amandeeepkour607@gmail.com