Data Mining - Evaluation

Data Warehouse

Data warehouse exibits following characteristics to support management's decision making process.:

- **Subject Oriented** - The Data warehouse is subject oriented because it provide us the information around a subject rather the organization's ongoing operations. These subjects can be product, customers, suppliers, sales, revenue etc. The data warehouse does not focus on the ongoing operations rather it focuses on modelling and analysis of data for decision making.
- **Integrated** - Data Warehouse is constructed by integration of data from heterogeneous sources such as relational databases, flat files etc. This integration enhance the effective analysis of data.
- **Time Variant** - The Data in Data Warehouse is identified with a particular time period. The data in data warehouse provide information from historical point of view.
- **Non volatile** - Non volatile means that the previous data is not removed when new data is added to it. The data warehouse is kept separate from the operational database therefore frequent changes in operational database is not reflected in data warehouse.

Data Warehousing

Data Warehousing is the process of constructing and using the data warehouse. The data warehouse is constructed by integrating the data from multiple heterogeneous sources. This data warehouse supports analytical reporting, structured and/or ad hoc queries and decision making.

Data Warehousing involves data cleaning, data integration and data consolidations. Integrating Heterogeneous Databases To integrate heterogeneous databases we have the two approaches as follows:

- Query Driven Approach
- Update Driven Approach

Query Driven Approach

This is the traditional approach to integrate heterogeneous databases. This approach was used to build wrappers and integrators on the top of multiple heterogeneous databases. These integrators are also known as mediators.

PROCESS OF QUERY DRIVEN APPROACH

- when the query is issued to a client side, a metadata dictionary translate the query into the queries appropriate for the individual heterogeneous site involved.
- Now these queries are mapped and sent to the local query processor.
- The results from heterogeneous sites are integrated into a global answer set.

DISADVANTAGES

This approach has the following disadvantages:

- The Query Driven Approach needs complex integration and filtering processes.
- This approach is very inefficient.
- This approach is very expensive for frequent queries.

- This approach is also very expensive for queries that requires aggregations.

Update Driven Approach

We are provided with the alternative approach to traditional approach. Today's Data Warehouse system follows update driven approach rather than the traditional approach discussed earlier. In Update driven approach the information from multiple heterogeneous sources is integrated in advance and stored in a warehouse. This information is available for direct querying and analysis.

ADVANTAGES

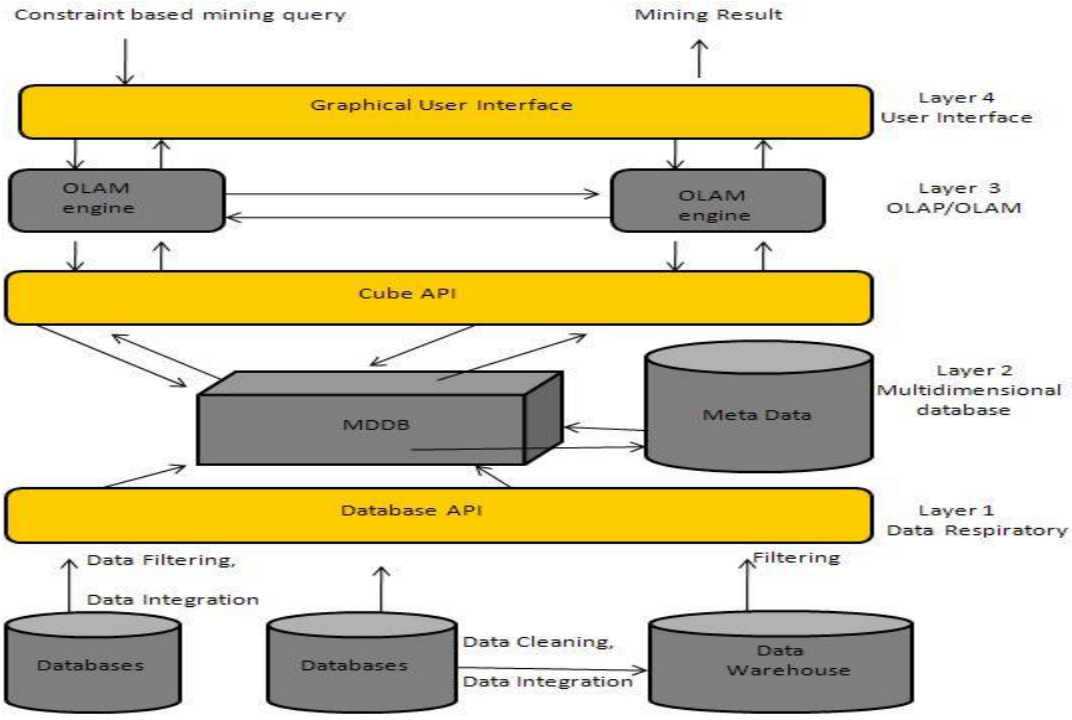
This approach has the following advantages:

- This approach provide high performance.
- The data are copied, processed, integrated, annotated, summarized and restructured in semantic data store in advance.

  Query processing does not require interface with the processing at local sources.

From Data Warehousing (OLAP) to Data Mining (OLAM)

Online Analytical Mining integrates with Online Analytical Processing with data mining and mining knowledge in multidimensional databases. Here is the diagram that shows integration of both OLAP and OLAM:



Importance of OLAM:

Here is the list of importance of OLAM:

- **High quality of data in data warehouses** - The data mining tools are required to work on integrated, consistent, and cleaned data. These steps are very costly in preprocessing of data. The data warehouse constructed by such preprocessing are valuable source of high quality data for OLAP and data mining as well.
- **Available information processing infrastructure surrounding data warehouses** - Information processing infrastructure refers to accessing, integration, consolidation, and transformation of multiple heterogeneous databases, web-accessing and service facilities, reporting and OLAP analysis tools.
- **OLAP-based exploratory data analysis** - Exploratory data analysis is required for effective data mining. OLAM provides facility for data mining on various sub set of data and at different level of abstraction.
- **Online selection of data mining functions** - Integrating OLAP with multiple data mining functions, on-line analytical mining provides users with the flexibility to select desired data mining functions and swap data mining tasks dynamically.

Source:

http://www.tutorialspoint.com/data_mining/dm_evaluation.htm