

Chapter - 3

Advanced Database Technologies

DATA WAREHOUSE AND DATA-MINING

Definition

The term data warehouse was coined with the definition of Inmon: "A warehouse is a subject-oriented, integrated, time variant and non-volatile collection of data in support of management's decision making process" .

Characteristics of DW

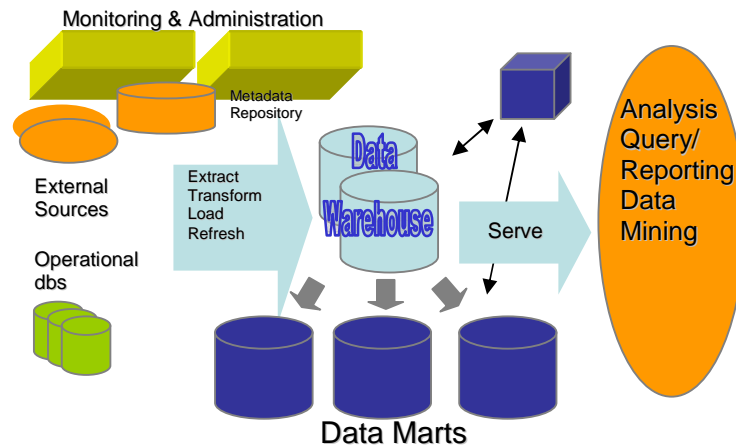
- **Subject-oriented :** Means that all relevant data about a subject is gathered and stored as a single set in a useful format. Information is presented according to specific subjects or areas of interest.
- **Time-variant :** Means that the data warehouse contains a history of the subject, as well as current information. It may be long-term data from five to ten years in contrast to the 30 to 60 day time..
- **Non-volatile :** Means stable information, Information is consistent; Data in the database is never over-written or deleted once committed.
- **Integrated :** Stored in a globally acceptable fashion with consistent naming conventions, measurements, encoding structures, and physical attributes, *though underlying operational systems store the data differently;*

Data Warehouse Architecture:

- **Warehouse database server:**
Which is almost always a relational DBMS; rarely flat files.
- **OLAP servers:-** Which may either be a ROLAP or MOLAP
 - *Relational OLAP (ROLAP):* **Extended relational DBMS that maps operations on multidimensional data to standard relational operations.**
 - *Multidimensional OLAP (MOLAP):* **Special purpose server that directly implements multidimensional data and operations.**
- **Clients:-**The Users or the client of Data warehouses are various Query and reporting tools, Analysis tools and Data mining tools (e.g., trend analysis, prediction)

Processors in Datawarehouse server

- **MPP**
 - **Massively parallel processing, a computer configuration that is able to use hundreds or thousands of CPUs simultaneously.**
- **SMP**
 - **Symmetric multi-processing is a computer configuration where many CPUs share a common operating system, main memory and disks. They can work on different parts of a problem at the same time.**



OLAP Vs OLTP

| | OLTP | Data Warehouse (DW) |
|-------------------|----------------------------------|------------------------------------|
| Type of Users | Clerk, IT Professional | Knowledge worker |
| Purpose | Day to day operations | Decision support |
| DB Model | Application-oriented (E-R based) | Subject-oriented (Star, snowflake) |
| Data | Current, Isolated | Historical, Consolidated |
| View | Detailed, Flat relational | Summarized, Multidimensional |
| Usage | Structured, Repetitive | Ad hoc |
| Unit of work | Short, Simple transaction | Complex query |
| Access Operations | Read/write | Read Mostly |
| No. of Users | Tens | Millions |
| Data base size | Thousands, 100 MB-GB | Hundreds, 100GB-TB |

Steps involved in creating a Datawarehouse

- **Data extraction**
- **Data cleaning**, also called *data cleansing* or *scrubbing*,
- **Data transformation**
- **Convert from legacy/host format to warehouse format**
- **Load**
- **Sort, summarize, consolidate, compute views, check integrity, build indexes, partition**
- **Refresh**
- **Propagate updates from sources to the warehouse**

Advantages

- Data warehouses enhance end-user access to a wide variety of data.
- Decision support system users can obtain specified trend reports, e.g. the item with the most sales in a particular area within the last two years.
- Data warehouses can be a significant enabler of commercial business applications.

Data Mining

- **Definition**

- Data mining, *the extraction of hidden predictive information from large databases*
- *The nontrivial extraction of implicit, previously unknown, and potentially useful information from data" and "the science of extracting useful information from large data sets or databases".*

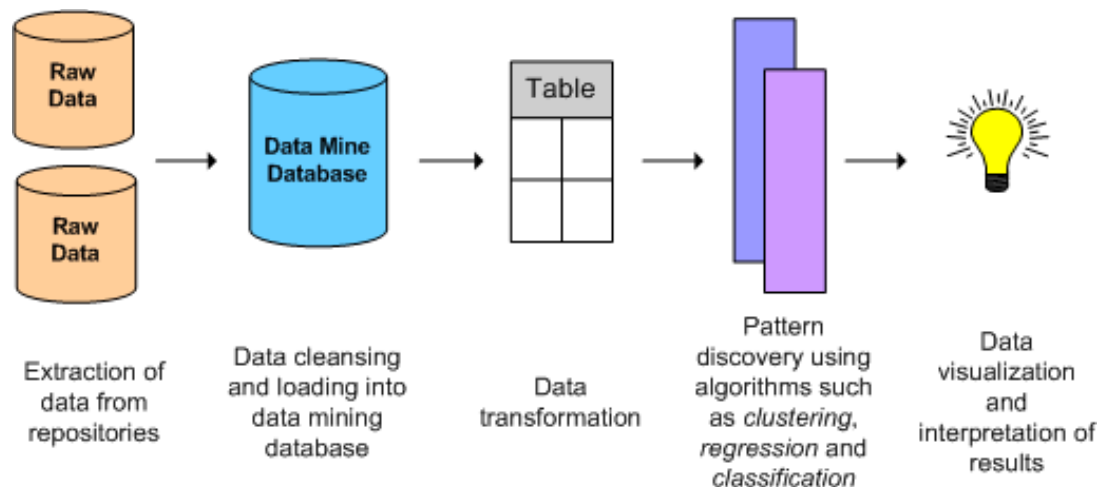
➤ **Techniques used in Data Mining: -**

Artificial neural networks
Genetic algorithms
Decision trees
Nearest neighbor method
Rule induction

Evolution of Datamining

| Evolutionary Step | Enabling Technologies | Characteristics |
|---|---|---|
| Data Collection (1960s) | Computers, tapes, disks | Retrospective, static data delivery |
| Data Access (1980s) | Relational databases (RDBMS), (SQL), ODBC | Retrospective, dynamic data delivery at record level |
| Data Warehousing & Decision Support (1990s) | On-line analytic processing (OLAP), multidimensional databases, data warehouses | Retrospective, dynamic data delivery at multiple levels |
| Data Mining (Emerging Today) | Advanced algorithms, multiprocessor comp, massive databases | Prospective, proactive information delivery |

Steps involved in Data mining process



* * *