Md Tarique Anwer

Lovely Professional University

Data Warehousing & Data Mining (Notes)

ECAP446

Table of Contents

[UNIT-01: Data Warehousing and Online Analytical Processing 2](#_Toc136756402)

[What is data warehousing? 2](#_Toc136756403)

# UNIT-01: Data Warehousing and Online Analytical Processing

## What is data warehousing?

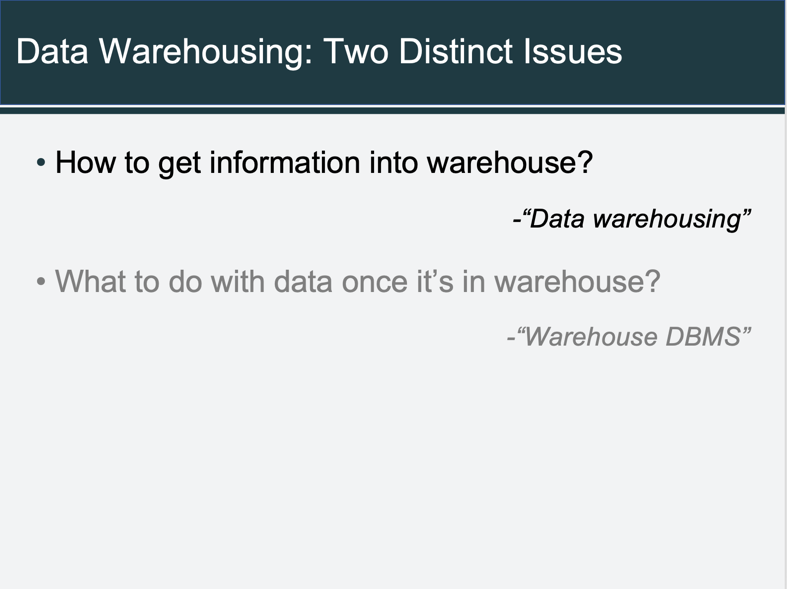
A data warehouse is a **subject-oriented**, **integrated**, **time variant** and **non-volatile** collection of data in support of management decision making process.

These four keywords - subject-oriented, integrated, time-variant, and nonvolatile distinguish data warehouses from other data repository systems, such as relational database systems, transaction processing systems, and file systems.

**Data warehouse properties**

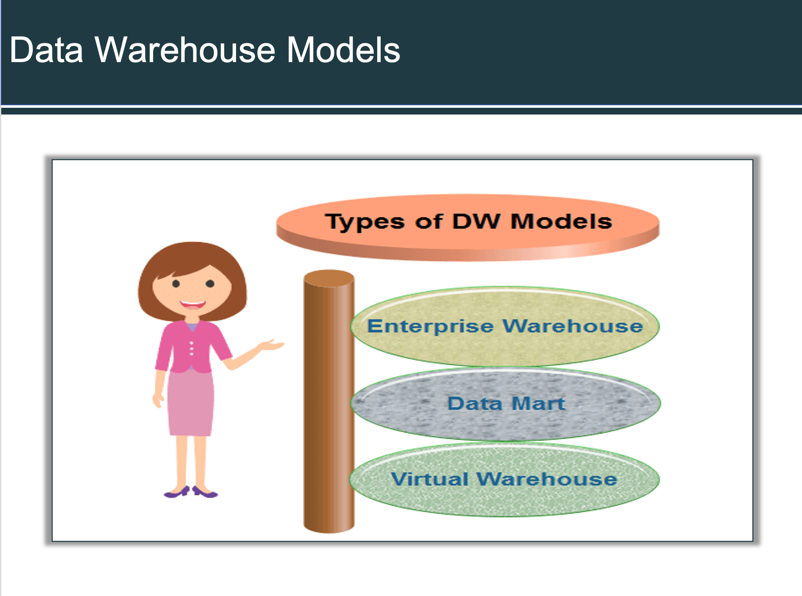
|  |  |  |  |
| --- | --- | --- | --- |
| **Subject-Oriented** | **Integrated** | **Time Variant** | **Non-Volatile** |
| Stored data according to target specific subjects.  **Example:** It may store data regarding total Sales, Number of Customers, etc. and not general data on everyday operations. | Data may be distributed across heterogeneous sources which have to be integrated.  **Example:** Sales data may be on RDB, Customer information on Flat files, etc. | Data are stored to provide information from an historic perspective.  **Example:** Last 5-10 years sales data | Operational Database and hence is not subject to frequent modification.  **Example:** It generally has 2 operations to perform on it:   * Loading of data * Access of data |

|  |  |
| --- | --- |
| **The Warehousing Approach** | **Heterogeneous Information Sources** |
| * Information integrated in advance. * Stored for direct querying and analysis. | * Different interfaces. * Different data representations. * Duplicate and inconsistent information |

**Features of a Warehouse**

* It is separate from Operational Database.
* Integrates data from heterogeneous systems.
* Stores HUGE amount of data, more historical than current data.
* Does not require data to be highly accurate.
* Queries are generally complex.
* Goal is to execute statistical queries and provide results which can influence decision making in favor of the Enterprise.
* These systems are thus called Online Analytical Processing Systems (OLAP).

**Need of a Separate Data Warehouse**

* + Use of OLAP query on OLTP system degrades system’s performance.
  + OLAP systems access historical data and not current volatile data while OLTP systems access current up-to-date data and do not need historical data.
  + An *Operational Database* is designed for known tasks like indexing and hashing using primary keys, searching for particular records, and many more.
  + Data Warehouse queries are often complex.
  + The computation of large data groups at summarized levels, and may require the use of special data organization, access, and implementation methods based on multidimensional views.

**Difference between OLAP & OLTP**

|  |  |
| --- | --- |
| **OLAP** | **OLTP** |
| Consists of historical data from various Databases. | Consists only operational current data. |
| It is subject oriented. | It is application oriented. |
| Used for Data Mining, Analytics, etc. | Used for business tasks. |
| The data is used in planning, problem solving and decision making. | The data is used to perform day to day fundamental operations. |
| It reveals a snapshot of present business tasks. | It provides a multi-dimensional view of different business tasks. |
| Relatively slow as more data involved. Queries may take hours. | Very Fast as the queries operate on 5% of the data. |
| It only need backup from time to time as compared to OLTP. | Backup and recovery process is maintained religiously. |
| This data is generally managed by CEO, MD, GM. | This data is managed by clerks, managers. |

Multi-dimensional data model

A data warehouse is based on multidimensional model which views data in the form of a data cube.