**Madiha Aimon Tappal**

[**madihaaimon@gmail.com**](mailto:madihaaimon@gmail.com)

**Data Engineering Batch – 1**

**Day – 2 Assignment**

**What is a Data Warehouse?**

* **Data Warehouse (DW)** is a Subject oriented, integrated, time variant, non-volatile collection of data in support of management’s system.
* A **Data Warehouse**is separate from DBMS, it stores a huge amount of data, which is typically collected from multiple heterogeneous sources like files, DBMS, etc. The goal is to produce statistical results that may help in decision-making. For example, a college might want to see quick different results, like how the placement of CS students has improved over the last 10 years, in terms of salaries, counts, etc.
* It is a collection of data designed to support management decision making by presenting a coherent picture of business conditions at a single point of time.
* Data Warehousing can be applied anywhere where we have a huge amount of data and we want to see statistical results that help in decision making.
* **E.g. Social Media Websites, Banking**

**Need for Data Warehouse**

* An ordinary Database can store MBs to GBs of data and that too for a specific purpose. For storing data of TB size, the storage shifted to the Data Warehouse.
* Besides this, a transactional database doesn’t offer itself to analytics. To effectively perform analytics, an organization keeps a central Data Warehouse to closely study its business by organizing, understanding, and using its historical data for making strategic decisions and analysing trends.
* Features of **Data Warehouse**: -
  + Subject-oriented
  + Integrated
  + Time-variant
  + Non-volatile
* Subject-oriented:
  + Data are organized according to the subject instead of application.
  + It mainly focuses on modelling and analysis of data for decision makers, not on daily operations or transaction processing.
* Integrated:
  + Constructed by integrating multiple, heterogeneous data sources like relational databases, flat files, on-line transaction records.
  + Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources.
* Time-variant:
  + The time horizon for the data warehouse is significantly longer than that of operational systems. i.e. provide information from a historical perspective (e.g., past 5-10 years).

* Non-volatile:
  + No updates are allowed. Once the data entered into the data warehouse, they are never removed.
  + The data in warehouse represent company’s history, the operational data representing near term history are always added to it.

### **Benefits of Data Warehouse**

* **Better business analytics:**Data warehouse plays an important role in every business to store and analysis of all the past data and records of the company. which can further increase the understanding or analysis of data for the company.
* **Faster Queries: The data** warehouse is designed to handle large queries that’s why it runs queries faster than the database.
* **Improved data Quality:**In the data warehouse the data you gathered from different sources is being stored and analysed it does not interfere with or add data by itself so your quality of data is maintained and if you get any issue regarding data quality then the data warehouse team will solve this.
* **Historical Insight:**The warehouse stores all your historical data which contains details about the business so that one can analyse it at any time and extract insights from it.

**Advantages of Data Warehousing**

* **Intelligent Decision-Making:**With centralized data in warehouses, decisions may be made more quickly and intelligently.
* **Business Intelligence:**Provides strong operational insights through business intelligence.
* **Historical Analysis:** Predictions and trend analysis are made easier by storing past data.
* **Data Quality:**Guarantees data quality and consistency for trustworthy reporting.
* **Scalability:** Capable of managing massive data volumes and expanding to meet changing requirements.
* **Effective Queries:**Fast and effective data retrieval is made possible by an optimized structure.
* **Cost reductions:** Data warehousing can result in cost savings over time by reducing data management procedures and increasing overall efficiency, even when there are setup costs initially.
* **Data security:** Data warehouses employ security protocols to safeguard confidential information, guaranteeing that only authorized personnel are granted access to certain data.

**Disadvantages of Data Warehousing**

* **Cost:**Building a data warehouse can be expensive, requiring significant investments in hardware, software, and personnel.
* **Complexity:**Data warehousing can be complex, and businesses may need to hire specialized personnel to manage the system.
* **Time-consuming:**Building a data warehouse can take a significant amount of time, requiring businesses to be patient and committed to the process.
* **Data integration challenges:**Data from different sources can be challenging to integrate, requiring significant effort to ensure consistency and accuracy.
* **Data security:**Data warehousing can pose data security risks, and businesses must take measures to protect sensitive data from unauthorized access or breaches.

**DATA WAREHOUSE: -**

* Data Warehouse is a database used for data reporting and analysis.
* The data stored in the warehouse

are uploaded from the operational systems (such as marketing, sales etc.)

* The data store contains two main types of data.

-Business Data

-Business Data Model

* The business data are extracted from operational database and from external sources which are relevant to the business (stock prices, market indicators, marketing information and competitor’s data).

**DSS data vs Operational data**

* The DSS data which is extracted from multiple sources differ from operational data in three main areas:

- time span,

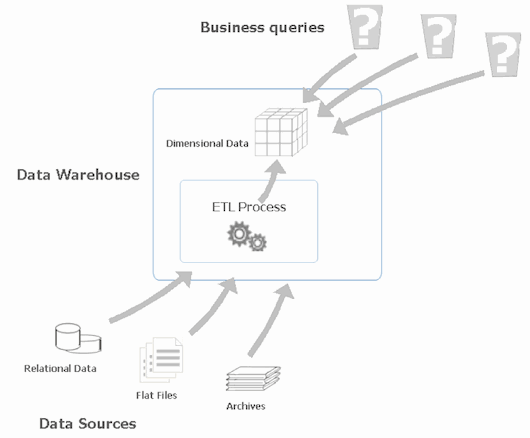
- granularity and

- dimensionality.

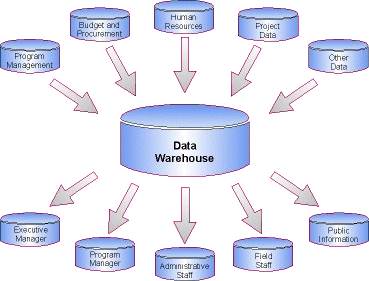
**A typical ETL (Extract Transform Load)-based data warehouse uses**

1. staging,
2. integration and
3. access layers (Data Marts)

to house its key functions.



* The data that arrived at data warehouse are first passed to Operational Data Store (ODS).
* Data is integrated from multiple sources for additional operations on the data.
* This integrated data is passed back to operational systems for decision-making.

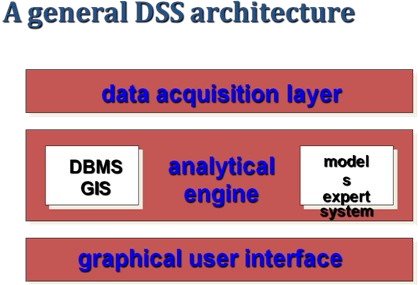


**Three kinds of data warehouse applications**

* Information processing
  + supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts and graphs.
* Analytical processing
  + multidimensional analysis of data warehouse data.
  + supports basic OLAP operations, slice-dice, drilling, pivoting.
* Data mining
  + knowledge discovery from hidden patterns.
  + supports associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools.

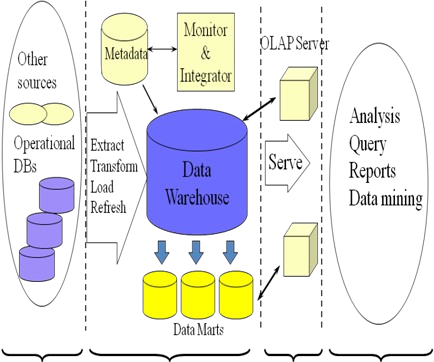
**Need for Decision Support System (DSS) in business**

* In a typical business environment with an increasing competitive market, we cannot ponder over decisions for too long.
* Hence the business needs managers with quick decision making.
* Therefore, to succeed in business today, any company needs information systems that can support the diverse information and decision-making needs.
* Decision Support Systems helps us to assess and resolve everyday business questions.
* It works by compiling useful information from a combination of raw data, documents, personal knowledge, or business models.

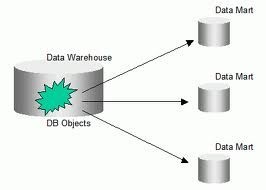


**DSS components**

* The DSS is usually composed of five main components: -
  + Data store component
  + Data extraction component
  + Data filtering component
  + End user query tool
  + End user presentation tool

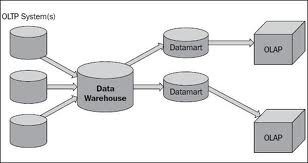


* The data in the data warehouse is stored in the form of Data marts.
* It allows the user to access the data in terms of a specific business line or team.



**Data mart: -**

* The data mart is a subset of the data warehouse that is usually oriented to a specific business line



**In order to implement DSS**

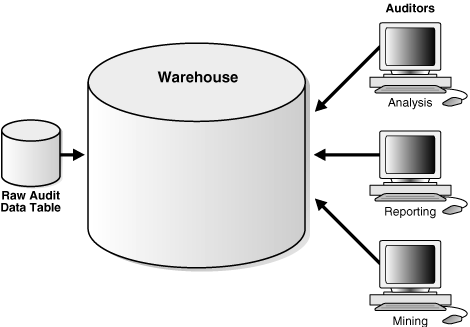
* **DSS architectural styles**
* OLTP (Online Transaction Processing)

- used by traditional operational systems (RDBMS).

* OLAP (Online Analytical Processing)
  + used by Data Warehouse.

**OLTP (Online Transaction Processing): -**

* **On-Line Transaction Processing (OLTP) System** refers to the system that manage transaction-oriented applications. These systems are designed to support on-line transaction and process query quickly on the Internet.
* **For example:** POS (point of sale) system of any supermarket is a OLTP System.
* Every industry in today’s world use OLTP system to record their transactional data. The main concern of OLTP systems is to enter, store and retrieve the data. They cover all day-to-day operations such as purchasing, manufacturing, payroll, accounting, etc. Of an organization. Such systems have large numbers of user which conduct short transaction. It supports simple database query so the response time of any user action is very fast.
* The data acquired through an OLTP system is stored in commercial RDBMS, which can be used by an OLAP System for data analytics and other business intelligence operations
* The **Operational Database** is one which is accessed by an Operational System to carry out regular operations of an organization.
* Operational Databases usually use an OLTP architecture which is optimized for faster transaction processing.
* OLTP Databases access the data in the form of operations like- Inserting, Deleting, and Updating data.
* **OLTP** is a methodology to provide end users with access to large amounts of data.
* It works in an intuitive and rapid manner to assist with deductions based on investigative reasoning.
* **OLTP** refers to a class of systems that facilitate and manage transaction-oriented applications, typically for data entry and retrieval transaction processing.



A typical OLTP Architecture

An **Automatic Teller Machine (ATM)**for a bank is an example of a commercial transaction processing application.

**Benefits of OLTP**

* Simplicity & Efficiency:

Reduced paper trails and the faster & more accurate forecasts for revenues and expenses are both examples of how OLTP makes things simpler for businesses.

* OLTP systems maintain data integrity and they also provide fast query processing in environments having multiple access.

**Pitfalls of OLTP:** -

* + OLTP requires instant update.
  + The data what we get from OLTP is not suitable for data analysis.
  + To perform one simple transaction even with the normalized structure, we need to query multiple tables by using joins.

**Advantages of an OLTP System: -**

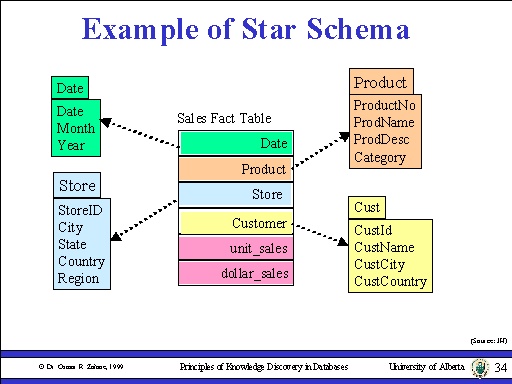
* OLTP Systems are user friendly and can be used by anyone having basic understanding
* It allows its user to perform operations like read, write and delete data quickly.
* It responds to its user actions immediately as it can process query very quickly.
* These systems are original source of the data.
* It helps to administrate and run fundamental business tasks
* It helps in widening customer base of an organization by simplifying individual processes

**OLAP (Online Analytical Processing): -**

* OLAP stands for **Online Analytical Processing** Server. It is a software technology that allows users to analyse information from multiple database systems at the same time. It is based on multidimensional data model and allows the user to query on multi-dimensional data
* OLAP is an approach to answer multi-dimensional analytical queries which also encompasses relational reporting and data mining.
* An OLAP cube is an array of data that is understood in terms of its 0 or more dimensions which enables the users to gain insight into their data in a fast, interactive, easy-to-use manner.
* For example, a company might wish to summarize financial data by product (Maruthi), by time-period (May 2006), by city (Mumbai) to compare actual budget expenses.

* Product, time, city and scenario (actual and budget) are the data's dimensions.
* OLAP data is typically stored in a Star Schema.

-which is a combination of dimensions and fact tables.

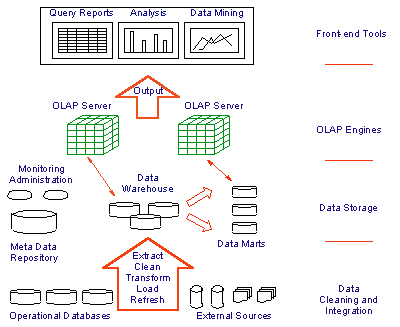


* act table consists of the measurement’s metrics or facts of a business process. It is located at the centre of a star schema surrounded by dimension tables.
* In the above example, it holding sales by item, by time, by branch and by location.
* Dimension tables contain descriptive attributes (or fields) which are typically textual fields or discrete numbers that behave like text.
* These attributes are designed to serve two critical purposes:

-query constraining/filtering

-query result set labelling.

**OLAP Architecture: -**



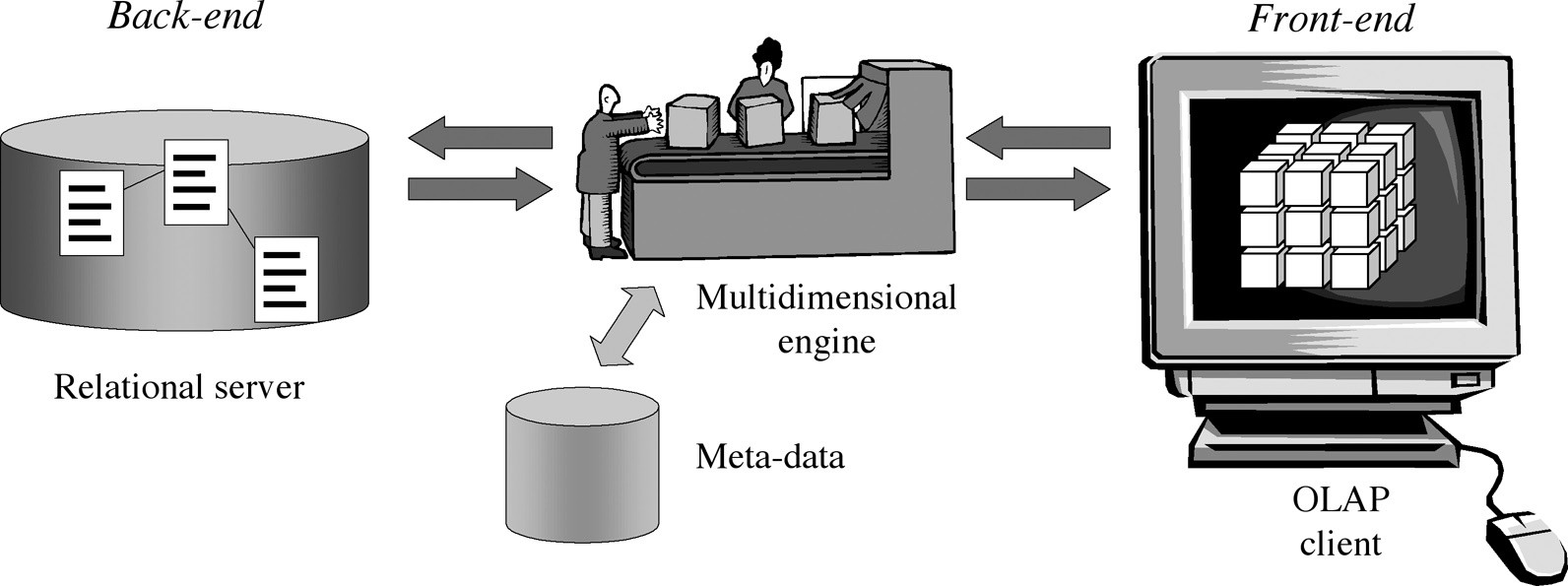
**OLAP Server: -**

* OLAP Server receives the data from data warehouse by which it represents the data in a user understandable format which actually supply analytical functionality for the DSS system.
* OLAP Server generally performs data analysis in two forms.

-ROLAP (Relational OLAP)

-MOLAP (Multi-dimensional OLAP)

* **ROLAP (Relational): -**
* It is a form of OLAP that performs dynamic multi- dimensional analysis of data stored in a relational database rather than in a multi-dimensional database (which is usually considered the OLAP standard).
* Data processing may take place within the database system, a mid-tier server, or the client.
* In two-tier architecture, the user submits a Structured Query Language (SQL) query to the database and receives back the requested data.



**Two-Tier Architecture of ROLAP**

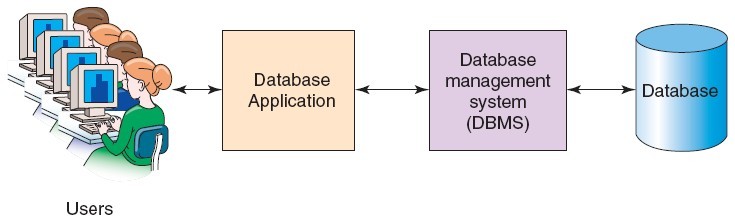
* **MOLAP (multi-dimensional): -**

* It is a form of OLAP that helps the user to “slice and dice” information, providing multi-dimensional analysis of data by putting data in a cube structure.

* Most MOLAP products use a multi-cube approach in which a series of small, dense, pre-calculated cubes make a hypercube.

## **What is RDBMS?**

* RDBMS stands for Relational Database Management Systems. It is basically a program that allows us to create, delete, and update a relational database. A Relational Database is a database system that stores and retrieves data in a tabular format organized in the form of rows and columns. It is a smaller subset of DBMS



## Characteristics of RDBMS

* Data must be stored in tabular form in DB file, that is, it should be organized in the form of rows and columns.
* Each row of table is called [record/tuple](https://www.geeksforgeeks.org/tuple-in-dbms/) . Collection of such records is known as the cardinality of the table.
* Each column of the table is called an attribute/field. Collection of such columns is called the arity of the table.
* No two records of the DB table can be same. Data duplicity is therefore avoided by using a candidate key. [Candidate Key](https://www.geeksforgeeks.org/candidate-key-in-dbms/)is a minimum set of attributes required to identify each record uniquely.
* Tables are related to each other with the help for foreign keys.
* Database tables also allow NULL values, that is if the values of any of the element of the table are not filled or are missing, it becomes a NULL value, which is not equivalent to zero. (NOTE: [Primary key](https://www.geeksforgeeks.org/primary-key-in-dbms/) cannot have a NULL value).

**Different types of Databases: -**

* Relational Databases
* Operational Databases
* Database Warehouses
* Distributed Databases
* End-User Database

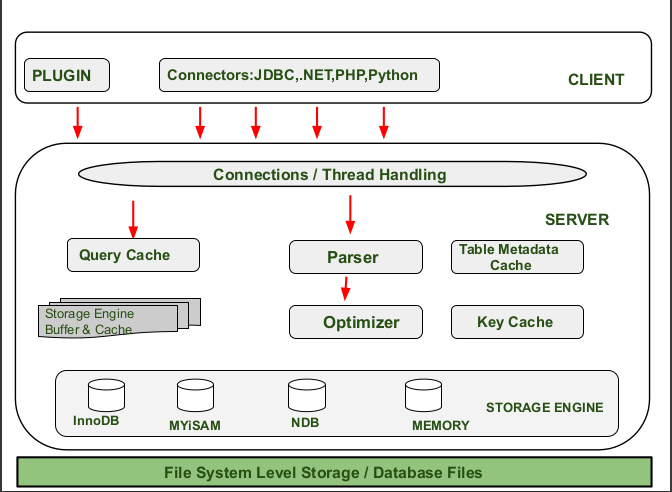
**Why MYSQL?**

* MySQL is the world's most popular open-source database software, with over 100 million copies of its software downloaded or distributed throughout its history.
* MySQL is RDBMS which runs a server, providing multi-user access to a number of databases.

* With its superior speed, reliability, and ease of use, MySQL has become the preferred choice for IT in all sectors or domains.

## **How Does MySQL Work?**

* MySQL is open-source and user-friendly. It creates a database to store and manipulate the data. To perform various operations users, make requests by typing specific statements. The server responds to the information from the user and displays it on the user side.



Why MySQL?

D

## **Characteristics of SQL: -**

* SQL is an “ANSI” and “ISO” standard computer language for creating and manipulating database.
* It allows the user to create, update, delete and retrieve data from a database.
* It is very simple and easy to learn.
* It is used specifically for relational databases.
* It works with database programs like DB2, Oracle, MS Access, etc.

**Features of MySQL: -**

* MySQL is written in C and C++ and its SQL parser is written in yacc (Yet Another Compiler Compiler).
* MySQL uses only just under 1 MB of RAM on your laptop while Oracle 9i installation uses 128 MB
* MySQL is great for database enabled websites while Oracle is made for enterprises.
* MySQL is portable.
* MySQL default port number is 3306.

## **Difference Between MYSQL and SQL: -**

| **MySQL** | **SQL** |
| --- | --- |
| * MySQL was owned by Oracle corporation. | * SQL is developed by Microsoft Corporation. |
| * MySQL is open source and accessible to any and everyone for free. | * SQL is not open to others for free. |
| * MySQL supports basic programming languages like C, C++, Python, Ruby, etc. | * SQL is in itself a programming language used for database systems. |
| * MySQL available only in the English language. | * SQL is available in different languages. |
| * MySQL doesn’t support user-defined functions and XML. | * SQL supports user-defined functions and XML. |

**Application of MySQL**: -

* MySQL used in E-Commerce websites.
* MySQL used in Data Warehousing.
* MySQL is used in the Login Application.