# LO 1 - Describe the components and function of database management systems.

* 1. – Explain various terms related to database management.

Discussion: Why do we use database?

* To store information
  + User data
  + Product data for ecommerce sites
  + Business-related data.

### Terminology

**Database**

* Self-describing collection of integrated tables (Kroenke)
* A shared collection of logically related data and its descriptions, designed to meet the needs of an organization (Conally / Begg)
* An organized collection of data, today typically in digital form. The data are typically organized to model relevant aspects of reality, in a way that supports processes requiring this information (Wikipedia)

**Metadata**

* Data about data
* System catalog or data dictionary.
* Self-describing nature of a database provides program-data independence.

**DBMS – Database Management System**

* Examples of such systems include: Access, Oracle, SQLite, MySQL, SQL Server, DB2, PostgreSQL (postgres)
* A set of programs used to define, administer, and process the database and its applications.
* The software that manages and controls access to the database (Connolly/Begg)
* Translates the user’s data requests to the physical data storage.

**RDBMS -** Relational Database Management System

* DBMS that organizes data using relations (tables)
* Other DBMS techniques include (<http://en.wikipedia.org/wiki/Database> AND <http://en.wikipedia.org/wiki/Data_model>):
  + Flat Files
  + Hierarchical
  + Network databases
  + Object-relational

**Schema** – The entire structure of the database including all Tables, Attributes, Relationships, etc.

**Relation** – the definition of a table with columns (attributes) and rows (tuples) but not the data itself.

Example: Author Table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **auID** | **auLName** | **auFName** | **Phone** | **Address** | **City** |
|  |  |  |  |  |  |

**Row, record, entry, tuple** –

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **auID** | **auLName** | **auFName** | **Phone** | **Address** | **City** |
| 001 | Smith | John | 333-3333 | Box 123 | Saskatoon |
| 002 | Brown | Jane | 444-4444 | Box 321 | Saskatoon |

**Column, field, header, attribute** –

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **auID** | **auLName** | **auFName** | **Phone** | **Address** | **City** |

**Constraint** – restrictions on data in a table.

**Domain** –

* the datatype (example: char, varchar, number, date),
* the size of the datatype (example: char(10)),
* whether the value is unique,
* whether the value exists in another table,
* whether the value has other constraints like business rules (**Check Constraint**). For example: An attribute can only be assigned a value between 0 and 9999.

**Superkey** –

**Candidate Key** –

**Primary Key** –

**Foreign Key** –

**Composite Key** –

**Unique Key** –

**Surrogate Key** –

**SQL** –

**Components of SQL**

* **DDL** –
* **DML** –
* **DCL** –

**4GL** –

**Transaction –**

**Data mining** –

**Big data** –

## 1.2 Discuss the historical development of DBMSs and the relational model and other database models

http://www.computerhistory.org/revolution/memory-storage/8/265/2207

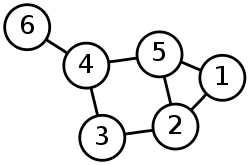
**File-based systems**

**Limitations of the File-Based Approach**

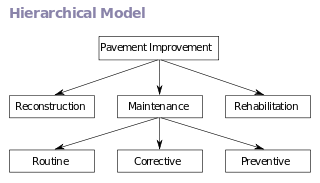
* Separation and isolation of data
* Duplication of data
* Data dependence
* Incompatible file formats
* Fixed queries

To solve the limitations of filebased systems the Database concept started in the 1960s.

The first generation of database systems were navigational , where applications accessed data by following pointers from one record to another. Ie. Node6.Node4.Node5.Node1

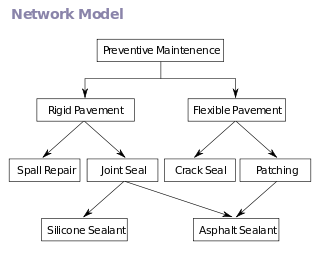


There were two main types of navigational databases, the network model (Codasyl a consortium model first proposed by Charles Bachman) and the hierarchical (IBM IMS system) model.

**The hierarchical model**



**The network model**





These systems were used until the 1980s where they were eventually replaced with relational databases. There was a resurgence of the hierarchical data systems in the late 1990s with the introduction of XML based systems.

**Relational Database Model**

The relational database model was proposed in 1970 by Edgar Codd. This model departed from the navigational model.

Relational database Models:



>

*Advantages of DBMS*



*Disadvantages*



**Object Oriented Database**

>

Potential advantages

Disadvantages

**XML Database**

>

**NoSQL database**

NoSQL (Not Only SQL) is the current trend mostly being utilized by the web development world. NoSQL is used to describe the new datastores that are non-relational. There are many different ways of classifying NoSQL databases such as by data model or feature. The 4 basic categories are:

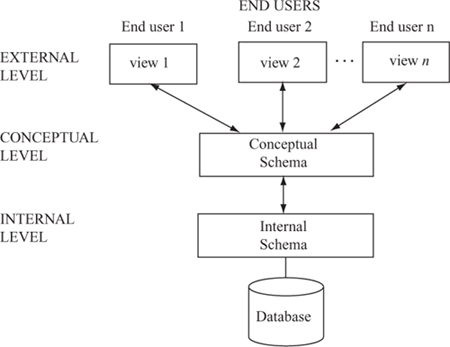
* Column –
* Document –
* Key-value –
* Graph –

## Diagram the architecture of a DBMS

When talking about DBMS architecture, we are referring to how the database is viewed by users. Abstract views of the data are provided to the users. These views hide how the data is stored so that the user does not need to worry about where data is located or how it’s stored.

The architecture is comprised of 3 different levels: external, conceptual and internal.

* External:   
  >
* Conceptual:   
  >
* Internal (Physical Level):   
    
  >



DBAs may have certain details

Developers/DBAs

End Users

Major components of DBMS architecture:

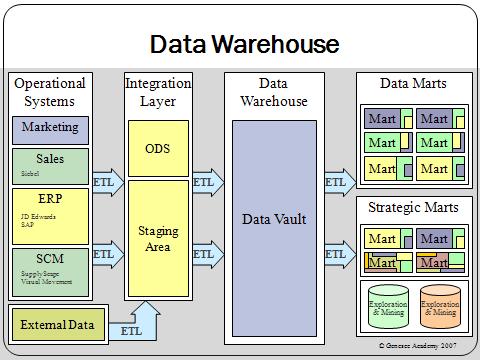
* **DBMS External Interfaces** –
* **Database Language Engines (or processors)** –
* **Query Optimizer** –
* **Database Engine** –
* **Storage Engine** –
* **Transaction Engine** –

## Discuss data warehouses and multidimensional data structures

### Data Warehouse

Data warehouse (DW) –

The data warehouse database tends to be:

1. - organized around subjects (customers, products, sales) instead of major functional areas (invoicing, stock control, product sales)
2. – integrating multiple data sources to be consistent.
3. – data represents a series of snapshots and is only valid at some point in time or some interval. (example: the change in price of a product over time)
4. – data is not added in real time but instead refreshed on a regular basis. Data is added as a supplement and does NOT replace old data. [](http://upload.wikimedia.org/wikipedia/commons/4/46/Data_warehouse_overview.JPG)

The **Extract, Transform and Load (ETL)** process:

These pieces of information are then stored in the Marts (Data and Strategic). Marts are used to improve performance and ease of use within that specific area. They are the access layer for providing data to the users.

Data mart –

Strategic mart –

The data is often grouped into hierarchical groups (trees) called dimensions and also into facts and aggregate facts. The combination of facts and dimensions is sometimes called a star schema.

There are two main approaches to storing data in a data warehouse, the dimensional approach (Star Schema) and the normalized approach.

In a dimensional approach, transaction data are partitioned into "facts", which are generally numeric transaction data, and "dimensions", which are the reference information that gives context to the facts.

Example: A sales transaction can be broken up into

1. facts:
2. and dimensions:

The main disadvantages of the dimensional approach are:

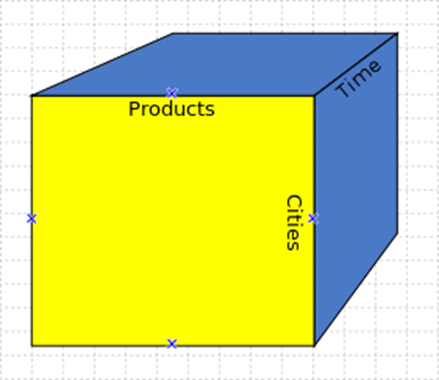
1.

2.

It should be noted that both normalized – and dimensional models can be represented in entity-relationship diagrams as both contain joined relational tables. The difference between the two models is the degree of normalization.

### Multidimensional Data Structures

In Relational databases, tables are only able to provide a 2 dimensional view of data. The multidimensional structure is similar to the relational model. The dimensions of the cube-like model have data relating to elements in each cell. This structure gives a spreadsheet-like view of data. This structure is easy to maintain because records are stored as fundamental attributes—in the same way they are viewed—and the structure is easy to understand. Its high performance has made it the most popular database structure when it comes to enabling online analytical processing (OLAP).



#### OnLine Analytical Processing

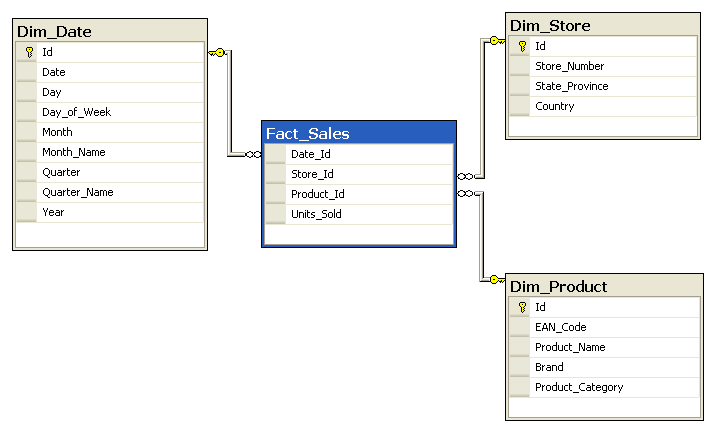
A cube can be thought of as a generalization of a two-dimensional spreadsheet. For example a company might wish to summarize financial data by product, by time-period, by city to compare actual and budget expenses. Product, time, city and scenario (actual and budget) are the data's dimensions.

Cube is a shortcut for multidimensional dataset, given that data can have an arbitrary number of dimensions. The term hypercube is sometimes used, especially for data with more than three dimensions.

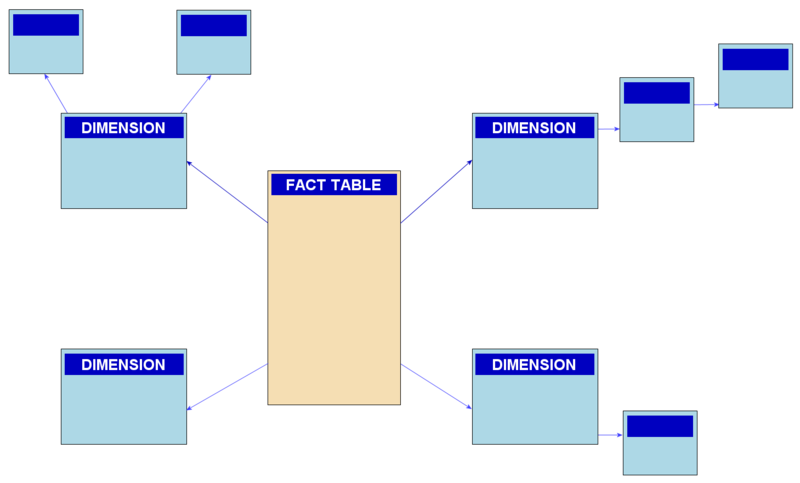
Each cell of the cube holds a number that represents some measure of the business, such as sales, profits, expenses, budget and forecast.

OLAP data is typically stored in a star schema or snowflake schema in a relational data warehouse or in a special-purpose data management system. Measures are derived from the records in the fact table and dimensions are derived from the dimension tables.

**Star Schema**



**Snowflake Schema**

[](http://upload.wikimedia.org/wikipedia/commons/b/b2/Snowflake-schema.png)

**Review for LO1:**

Work on review questions document.

Optional Resources and Extra Information:

Data mining – analyzing collected data (often stored in a data warehouse) for patterns.

Data mining video (Viewer discretion advised - this video uses language that some may find offensive. It also shows a dramatization of a pregnant woman’s water breaking). <https://www.youtube.com/watch?v=f2Kji24833Y>

Big data – collection of large and complex data sets to which data mining is usually applied.

Ted talk 22 minutes on big data: <https://www.youtube.com/watch?v=Zr02fMBfuRA>