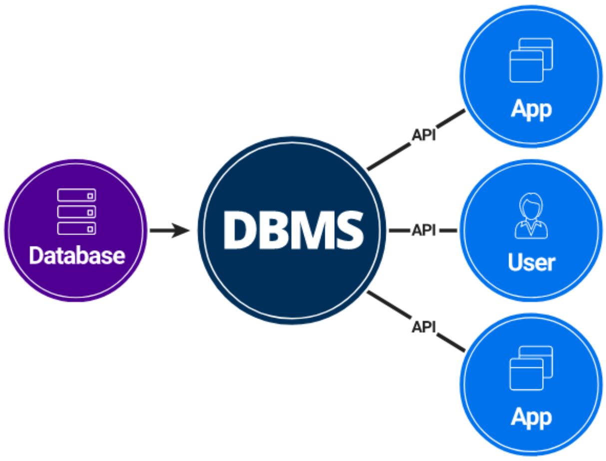
Catalog

LEC-1: Introduction to DBMS

1. **What is Data?**
    a. Data is a collection of raw, unorganized facts and details like text, observations, figures, symbols, and descriptions of things etc.
    In other words, **data does not carry any specific purpose and has no significance by itself.** Moreover, data is measured in terms of bits and bytes – which are basic units of information in the context of computer storage and processing.
    b. Data can be recorded and doesn't have any meaning unless processed.
2. **Types of Data**
    a. **Quantitative**
        i. Numerical form
        ii. Weight, volume, cost of an item.
    b. **Qualitative**
        i. Descriptive, but not numerical.
        ii. Name, gender, hair color of a person.
3. What is **Information**?
    a. Info. Is **processed, organized, and structured data.**
    b. It provides **context of the data and enables decision making.**
    c. Processed data that make **sense** to us.
    d. Information is extracted from the data, by **analyzing and interpreting** pieces of data.
    e. E.g., you have data of all the people living in your locality, its Data, when you analyze and interpret the data and come to some conclusion that:
        i. There are 100 senior citizens.
        ii. The sex ratio is 1.1.
        iii. Newborn babies are 100.
        These are information.
4. **Data vs Information**
    a. Data is a collection of facts, while information puts those facts into context.
    b. While data is raw and unorganized, information is organized.
    c. Data points are individual and sometimes unrelated. Information maps out that data to provide a big-picture view of how it all fits together.
    d. Data, on its own, is meaningless. When it's analyzed and interpreted, it becomes meaningful information.
    e. Data does not depend on information; however, information depends on data.
    f. Data typically comes in the form of graphs, numbers, figures, or statistics. Information is typically presented through words, language, thoughts, and ideas.
    g. Data isn't sufficient for **decision-making**, but you can make decisions based on information.
5. What is **Database**?
    a. Database is an electronic place/system where data is stored in a way that it can be **easily accessed, managed, and updated.**
    b. To make real use Data, we need **Database management systems. (DBMS)**
6. What is **DBMS**?
    a. A database-management system (DBMS) is a collection of **interrelated data** and **a set of programs to access those data**. The collection of data, usually referred to as the **database**, contains information relevant to an enterprise. The primary goal of a DBMS is to provide a way to **store and retrieve database information** that is both convenient and efficient.
    b. A DBMS is the database itself, along with all the software and functionality. It is used to perform different operations, like **addition**, **access**, **updating**, and **deletion** of the data.



7.

8. DBMS vs File Systems

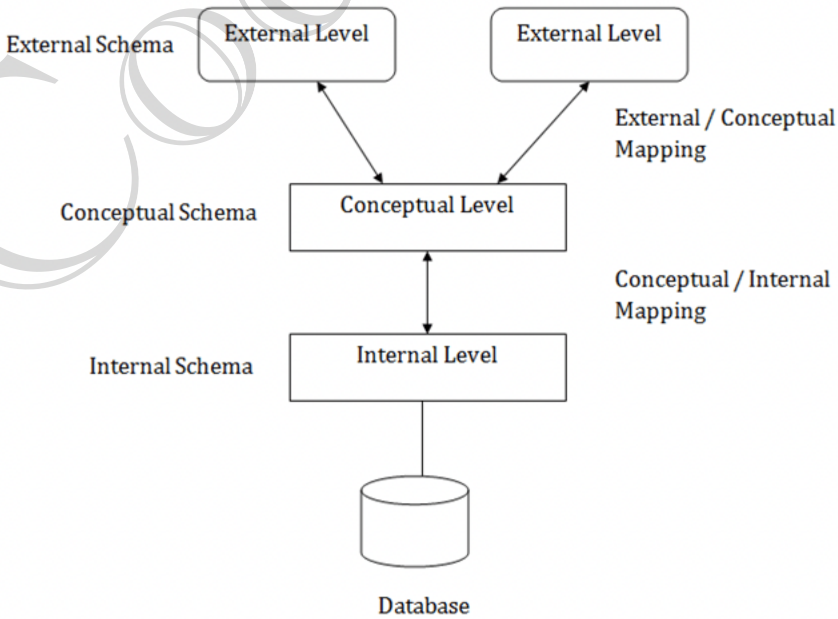
a. **File-processing systems** has major **disadvantages**.

i. Data Redundancy and inconsistency
ii. Difficulty in accessing data
iii. Data isolation
iv. Integrity problems
v. Atomicity problems
vi. Concurrent-access anomalies
vii. Security problems

b. Above 7 are also the **Advantages of DBMS** (answer to "**Why to use DBMS?**")

LEC-2: DBMS Architecture

1. **View of Data (Three Schema Architecture)**
   a. The major purpose of DBMS is to provide users with an **abstract view** of the data. That is, the system hides certain details of how the data is stored and maintained.
   b. To simplify user interaction with the system, abstraction is applied through several levels of abstraction.
   c. The **main objective** of three level architecture is to enable multiple users to access the same data with a personalized view while storing the underlying data only once
   d. **Physical level / Internal level**
      i. The lowest level of abstraction describes how the data are stored.
      ii. Low-level data structures used.
      iii. It has **Physical schema** which describes physical storage structure of DB.
      iv. Talks about: Storage allocation (N-ary tree etc), Data compression & encryption etc.
      v. **Goal**: We must define algorithms that allow efficient access to data.
   e. **Logical level / Conceptual level:**
      i. The **conceptual schema** describes the design of a database at the conceptual level, describes ***what*** data are stored in DB, and what ***relationships*** exist among those data.
      ii. User at logical level does not need to be aware about physical-level structures.
      iii. **DBA**, who must decide what information to keep in the DB use the logical level of abstraction.
      iv. **Goal**: ease to use.
   f. **View level / External level:**
      i. Highest level of abstraction aims to simplify users' interaction with the system by providing different view to different **end**-user.
      ii. Each **view schema** describes the database part that a particular user group is interested and hides the remaining database from that user group.
      iii. At the external level, a database contains several schemas that sometimes called as **subschema**. The subschema is used to describe the different view of the database.
      iv. At views also provide a **security** mechanism to prevent users from accessing certain parts of DB.



   g.

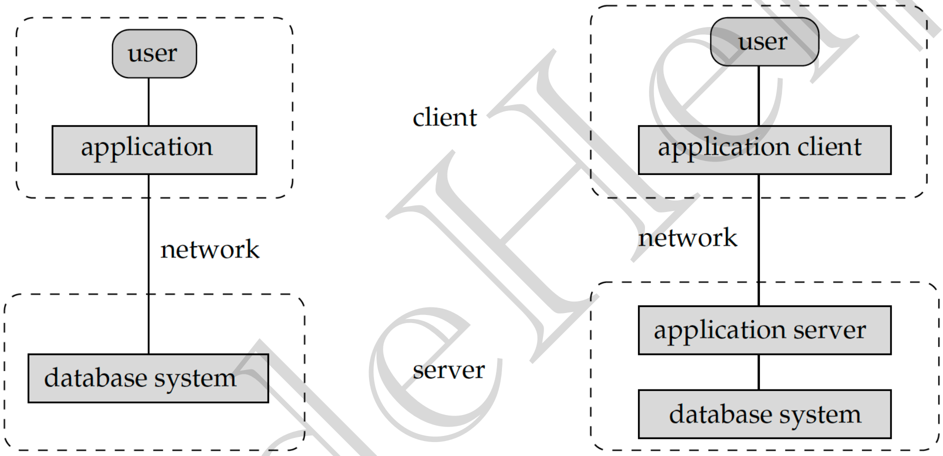
2. **Instances and Schemas**
   a. The collection of information stored in the DB at a particular moment is called an **instance** of DB.

b. The overall design of the DB is called the DB **schema**.
c. Schema is **structural** description of data. Schema **doesn't change frequently**. Data may change frequently.
d. **DB schema** corresponds to the variable declarations (along with type) in a program.
e. We have 3 types of **Schemas**: **Physical, Logical**, several **view schemas** called subschemas.
f. Logical schema is most **important** in terms of its effect on application programs, as programmers construct apps by using logical schema.
g. **Physical data independence**, physical schema change should not affect logical schema/application programs.

3. **Data Models:**
    a. Provides a way to describe the **design** of a DB at **logical level.**
    b. Underlying the structure of the DB is the Data Model; a collection of conceptual tools for describing **data, data relationships, data semantics & consistency constraints**.
    c. E.g., **ER** model, **Relational** Model, **object-oriented** model, **object-relational** data model etc.
4. **Database Languages**:
    a. **Data definition language (DDL)** to specify the database schema.
    b. **Data manipulation language (DML)** to express database queries and updates.
    c. **Practically**, both language features are present in a single DB language, e.g., SQL language.
    d. DDL
        i. We specify consistency constraints, which must be checked, every time DB is updated.
    e. DML
        i. Data manipulation involves
            1. **Retrieval** of information stored in DB.
            2. **Insertion** of new information into DB.
            3. **Deletion** of information from the DB.
            4. **Updating** existing information stored in DB.
        ii. **Query language**, a part of DML to specify statement requesting the retrieval of information.
5. **How is Database accessed from Application programs?**
    a. Apps (written in host languages, C/C++, Java) interacts with DB.
    b. E.g., Banking system's module generating payrolls access DB by executing DML statements from the host language.
    c. API is provided to send DML/DDL statements to DB and retrieve the results.
        i. Open Database Connectivity (**ODBC**), Microsoft "C".
        ii. Java Database Connectivity (**JDBC**), Java.
6. **Database Administrator (DBA)**
    a. A person who has **central control** of both the data and the programs that access those data.
    b. **Functions** of DBA
        i. Schema Definition
        ii. Storage structure and access methods.
        iii. Schema and physical organization modifications.
        iv. Authorization control.
        v. Routine maintenance
            1. Periodic backups.
            2. Security patches.
            3. Any upgrades.
7. **DBMS Application Architectures**: Client machines, on which remote DB users work, and server machines on which DB system runs.
    a. **T1** Architecture
        i. The client, server & DB all present on the same machine.

b. **T2** Architecture
   i. App is partitioned into 2-components.
   ii. Client machine, which invokes DB system functionality at server end through query language statements.
   iii. API standards like **ODBC** & **JDBC** are used to interact between client and server.

c. **T3** Architecture
   i. App is partitioned into 3 logical components.
   ii. Client machine is just a frontend and doesn't contain any direct DB calls.
   iii. Client machine communicates with App server, and App server communicated with DB system to access data.
   iv. **Business** logic, what action to take at that condition is in App server itself.
   v. T3 architecture are best for **WWW** Applications.
   vi. **Advantages**:
      1. **Scalability** due to distributed application servers.
      2. **Data integrity**, App server acts as a middle layer between client and DB, which minimize the chances of data corruption.
      3. **Security**, client can't directly access DB, hence it is more secure.



a. two-tier architecture

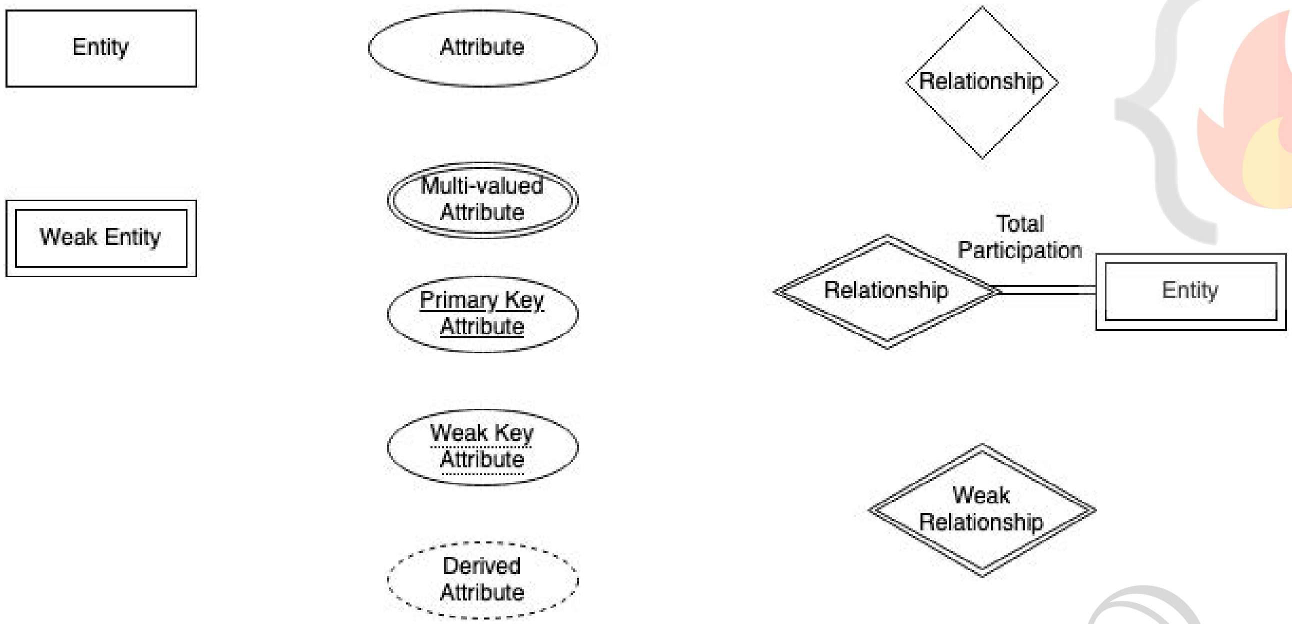
b. three-tier architecture

LEC-3: Entity-Relationship Model

1. **Data Model**: Collection of conceptual tools for **describing data, data relationships, data semantics, and consistency constraints.**
2. **ER Model**
   1. It is a high level data model based on a perception of a **real** world that consists of a collection of basic objects, called **entities** and of **relationships** among these objects.
   2. Graphical representation of ER Model is **ER diagram,** which acts as a **blueprint** of DB.
3. **Entity**: An Entity is a "**thing**" or "**object**" in the real world that is **distinguishable** from all other objects.
   1. It has **physical existence**.
   2. Each student in a college is an entity.
   3. Entity can be **uniquely** identified. (By a primary attribute, aka Primary Key)
   4. **Strong Entity**: Can be uniquely identified.
   5. **Weak Entity**: Can't be uniquely identified., depends on some other strong entity.
      1. It doesn't have sufficient attributes, to select a uniquely identifiable attribute.
      2. Loan -> Strong Entity, Payment -> Weak, as instalments are sequential number counter can be generated separate for each loan.
      3. **Weak entity depends on strong entity for existence.**
4. **Entity set**
   1. It is a set of entities of the **same** type that share the **same** properties, or attributes.
   2. E.g., Student is an entity set.
   3. E.g., Customer of a bank
5. **Attributes**
   1. An entity is represented by a set of attributes.
   2. Each entity has a value for each of its attributes.
   3. For each attribute, there is a set of **permitted values**, called the **domain**, or **value** set, of that attribute.
   4. E.g., Student Entity has following attributes
      A. Student_ID
      B. Name
      C. Standard
      D. Course
      E. Batch
      F. Contact number
      G. Address
   5. **Types of Attributes**
      1. **Simple**
         1. Attributes which can't be divided further.
         2. E.g., Customer's account number in a bank, Student's Roll number etc.
      2. **Composite**
         1. Can be divided into subparts (that is, other attributes).
         2. E.g., Name of a person, can be divided into first-name, middle-name, last-name.
         3. If user wants to refer to an entire attribute or to only a component of the attribute.
         4. Address can also be divided, street, city, state, PIN code.
      3. **Single-valued**
         1. Only one value attribute.
         2. e.g., Student ID, loan-number for a loan.
      4. **Multi-valued**
         1. Attribute having more than one value.
         2. e.g., phone-number, nominee-name on some insurance, dependent-name etc.
         3. Limit constraint may be applied, upper or lower limits.
      5. **Derived**
         1. Value of this type of attribute can be derived from the value of other related attributes.

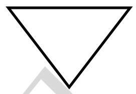
        2. e.g., Age, loan-age, membership-period etc.
    6. **NULL Value**
        1. An attribute takes a null value when an entity does not have a value for it.
        2. It may indicate "not applicable", value doesn't exist. e.g., person having no middle-name
        3. It may indicate "unknown".
            1. Unknown can indicate missing entry, e.g., name value of a customer is NULL, means it is missing as name must have some value.
            2. Not known, salary attribute value of an employee is null, means it is not known yet.
6. **Relationships**
    1. **Association** among two or more entities.
    2. e.g., Person has vehicle, Parent has Child, Customer borrow loan etc.
    3. **Strong Relationship**, between two independent entities.
    4. **Weak Relationship**, between weak entity and its owner/strong entity.
        1. e.g., Loan <instalment-payments> Payment.
    5. **Degree of Relationship**
        1. Number of entities participating in a relationship.
        2. **Unary**, Only one entity participates. e.g., Employee manages employee.
        3. **Binary**, two entities participates. e.g., Student takes Course.
        4. **Ternary** relationship, three entities participates. E.g, Employee works-on branch, employee works-on job.
        5. Binary are **common**.
7. **Relationships Constraints**
        1. **Mapping Cardinality / Cardinality Ratio**
            1. Number of entities to which another entity can be associated via a relationship.
            2. **One to one,** Entity in A associates with at most one entity in B, where A & B are entity sets. And an entity of B is associated with at most one entity of A.
                1. E.g., Citizen has Aadhar Card.
            3. **One to many,** Entity in A associated with N entity in B. While entity in B is associated with at most one entity in A.
                1. e.g., Citizen has Vehicle.
            4. **Many to one,** Entity in A associated with at most one entity in B. While entity in B can be associated with N entity in A.
                1. e.g., Course taken by Professor.
            5. **Many to many,** Entity in A associated with N entity in B. While entity in B also associated with N entity in A.
                1. Customer buys product.
                2. Student attend course.
        2. **Participation Constraints**
            1. Aka, **Minimum cardinality constraint.**
            2. **Types**, Partial & Total Participation.
            3. **Partial Participation,** not all entities are involved in the relationship instance.
            4. **Total Participation**, each entity must be involved in at least one relationship instance.
            5. e.g., Customer borrow loan, loan has total participation as it can't exist without customer entity. And customer has partial participation.
            6. **Weak entity has total participation constraint, but strong may not have total.**
8. **ER Notations**

Symbols used in ER Diagram



LEC-4: Extended ER Features

1. **Basic ER Features** studied in the **LEC-3,** can be used to model most DB features but when complexity increases, it is better to use some Extended ER features to model the DB Schema.
2. **Specialisation**
    1. In ER model, we may require to subgroup an entity set into other entity sets that are distinct in some way with other entity sets.
    2. **Specialisation** is **splitting** up the entity set into further **sub entity sets** on the basis of their **functionalities**, **specialities** and **features**.
    3. It is a **Top-Down** approach.
    4. e.g., **Person** entity set can be divided into **customer**, **student**, **employee**. Person is **superclass** and other specialised entity sets are **subclasses**.
        1. We have **"is-a"** relationship between superclass and subclass.
        2. Depicted by **triangle** component.



    5. **Why** Specialisation?
        1. Certain attributes may only be applicable to a few entities of the parent entity set.
        2. DB designer can show the distinctive features of the sub entities.
        3. To group such entities we apply Specialisation, to overall **refine** the DB blueprint.
3. **Generalisation**
    1. It is just a **reverse** of Specialisation.
    2. DB Designer, may encounter certain properties of two entities are overlapping. Designer may consider to make a new generalised entity set. That generalised entity set will be a super class.
    3. **"is-a"** relationship is present between subclass and super class.
    4. e.g., **Car**, **Jeep** and **Bus** all have some common attributes, to avoid data repetition for the common attributes. DB designer may consider to Generalise to a new entity set "**Vehicle**".
    5. It is a **Bottom-up** approach.
    6. **Why** Generalisation?
        1. Makes DB more **refined** and **simpler**.
        2. Common attributes are not **repeated**.
4. **Attribute Inheritance**
    1. **Both** Specialisation and Generalisation, has attribute inheritance.
    2. The attributes of higher level entity sets are inherited by lower level entity sets.
    3. E.g., **Customer & Employee** inherit the attributes of **Person**.
5. **Participation Inheritance**
    1. If a parent entity set participates in a relationship then its child entity sets will also participate in that relationship.
6. **Aggregation**
    1. **How to show relationships among relationships?** - Aggregation is the technique.
    2. **Abstraction** is applied to treat relationships as higher-level entities. We can call it Abstract entity.
    3. **Avoid redundancy** by aggregating relationship as an entity set itself.

LEC-7: Relational Model

1. Relational Model (RM) organises the data in the form of **relations** (**tables**).
2. A relational DB consists of **collection of tables**, each of which is assigned a **unique name.**
3. A **row** in a table represents a relationship among a set of values, and table is collection of such relationships.
4. **Tuple**: A single row of the table representing a single data point / a unique record.
5. **Columns**: represents the attributes of the relation. Each attribute, there is a permitted value, called **domain of the attribute.**
6. **Relation Schema**: defines the design and structure of the relation, contains the name of the relation and all the columns/attributes.
7. Common RM based DBMS systems, aka RDBMS: Oracle, IBM, **MySQL**, MS Access.
8. **Degree of table:** number of attributes/columns in a given table/relation.
9. **Cardinality**: Total no. of tuples in a given relation.
10. **Relational Key**: Set of attributes which can uniquely identify an each tuple.
11. Important **properties** of a **Table** in Relational Model
    1. The name of relation is distinct among all other relation.
    2. The values have to be atomic. Can't be broken down further.
    3. The name of each attribute/column must be unique.
    4. Each tuple must be unique in a table.
    5. The sequence of row and column has no significance.
    6. Tables must follow integrity constraints - it helps to maintain data consistency across the tables.
12. **Relational Model Keys**
    1. Super Key (**SK**): Any P&C of attributes present in a table which can uniquely identify each tuple.
    2. Candidate Key (**CK**): minimum subset of super keys, which can uniquely identify each tuple. It contains no redundant attribute.
        1. **CK value shouldn't be NULL.**
    3. Primary Key (**PK**):
        1. Selected out of CK set, has the least no. of attributes.
    4. Alternate Key (**AK**)
        1. All CK except PK.
    5. Foreign Key (**FK**)
        1. It creates relation between two tables.
        2. A relation, say r1, may include among its attributes the PK of an other relation, say r2. This attribute is called FK from r1 referencing r2.
        3. The relation r1 is aka **Referencing (Child) relation** of the FK dependency, and r2 is called **Referenced (Parent) relation** of the FK.
        4. FK helps to cross reference between two different relations.
    6. **Composite** Key: PK formed using at least 2 attributes.
    7. **Compound** Key: PK which is formed using 2 FK.
    8. **Surrogate** Key:
        1. Synthetic PK.
        2. Generated automatically by DB, usually an integer value.
        3. May be used as PK.
13. **Integrity Constraints**
    1. CRUD Operations must be done with some integrity policy so that DB is always consistent.
    2. Introduced so that we do not accidentally corrupt the DB.
    3. **Domain Constraints**
        1. Restricts the value in the attribute of relation, specifies the Domain.
        2. Restrict the Data types of every attribute.
        3. E.g., We want to specify that the enrolment should happen for candidate birth year < 2002.
    4. **Entity Constraints**
        1. Every relation should have PK. PK != NULL.

5. **Referential Constraints**
    1. Specified between two relations & helps maintain consistency among tuples of two relations.
    2. It requires that the value appearing in specified attributes of any tuple in referencing relation also appear in the specified attributes of at least one tuple in the referenced relation.
    3. If FK in referencing table refers to PK of referenced table then every value of the FK in referencing table must be NULL or available in referenced table.
    4. FK must have the matching PK for its each value in the parent table or it must be NULL.
6. **Key Constraints:** The six types of key constraints present in the Database management system are:-
    1. NOT NULL: This constraint will restrict the user from not having a NULL value. It ensures that every element in the database has a value.
    2. UNIQUE: It helps us to ensure that all the values consisting in a column are different from each other.
    3. DEFAULT: it is used to set the default value to the column. The default value is added to the columns if no value is specified for them.
    4. CHECK: It is one of the integrity constraints in DBMS. It keeps the check that integrity of data is maintained before and after the completion of the CRUD.
    5. PRIMARY KEY: This is an attribute or set of attributes that can uniquely identify each entity in the entity set. The primary key must contain unique as well as not null values.
    6. FOREIGN KEY: Whenever there is some relationship between two entities, there must be some common attribute between them. This common attribute must be the primary key of an entity set and will become the foreign key of another entity set. This key will prevent every action which can result in loss of connection between tables.

LEC-8: Transform - ER Model to Relational Model

1. Both **ER**-Model and **Relational** Model are abstract logical representation of real world enterprises. Because the two models implies the similar design principles, **we can convert ER design into Relational design**.
2. Converting a DB representation from an ER diagram to a table format is the way we arrive at Relational DB-design from an ER diagram.
3. ER diagram **notations to relations**:
   1. **Strong Entity**
      1. Becomes an **individual table** with entity name, attributes becomes columns of the relation.
      2. Entity's Primary Key (PK) is used as Relation's PK.
      3. **FK** are added to establish relationships with other relations.
   2. **Weak Entity**
      1. A table is formed with all the attributes of the entity.
      2. PK of its corresponding Strong Entity will be added as **FK**.
      3. **PK** of the relation will be a composite PK, {FK + Partial discriminator Key}.
   3. **Single Values Attributes**
      1. Represented as **columns** directly in the tables/relations.
   4. **Composite Attributes**
      1. Handled by **creating a separate attribute** itself in the original relation for each composite attribute.
      2. e.g., **Address**: {street-name, house-no}, is a composite attribute in customer relation, we add address-street-name & address-house-name as new columns in the attribute and ignore "address" as an attribute.
   5. **Multivalued Attributes**
      1. **New tables** (named as original attribute name) are created for each multivalued attribute.
      2. PK of the entity is used as column **FK** in the new table.
      3. Multivalued attribute's similar name is added as a column to define multiple values.
      4. **PK** of the new table would be {FK + multivalued name}.
      5. e.g., For Strong entity **Employee**, **dependent-name** is a multivalued attribute.
         1. New table named dependent-name will be formed with columns emp-id, and dname.
         2. PK: {emp-id, name}
         3. FK: {emp-id}
   6. **Derived Attributes**: Not considered in the tables.
   7. **Generalisation**
      1. **Method-1:** Create a table for the higher level entity set. For each lower-level entity set, create a table that includes a column for each of the attributes of that entity set plus a column for each attribute of the primary key of the higher-level entity set.
         For e.g., Banking System generalisation of Account - saving & current.
         1. Table 1: account (<u>account-number</u>, balance)
         2. Table 2: savings-account (<u>account-number</u>, interest-rate, daily-withdrawal-limit)
         3. Table 3: current-account (<u>account-number</u>, overdraft-amount, per-transaction-charges)
      2. **Method-2:** An alternative representation is possible, if the generalisation is disjoint and complete—that is, if no entity is a member of two lower-level entity sets directly below a higher-level entity set, and if every entity in the higher level entity set is also a member of one of the lower-level entity sets. Here, do not create a table for the higher-level entity set. Instead, for each lower-level entity set, create a table that includes a column for each of the attributes of that entity set plus a column for each attribute of the higher-level entity sets.
         Tables would be:
         1. Table 1: savings-account (<u>account-number</u>, balance, interest-rate, daily-withdrawal-limit)
         2. Table 2: current-account (<u>account-number</u>, balance, overdraft-amount, per-transaction-charges)
      3. **<u>Drawbacks of Method-2:</u>** If the second method were used for an overlapping generalisation, some values such as balance would be stored twice unnecessarily. Similarly, if the generalisation were not complete—that is, if some accounts were neither savings nor current accounts—then such accounts could not be represented with the second method.
   8. **Aggregation**

1. Table of the relationship set is made.
2. Attributes includes primary keys of entity set and aggregation set's entities.
3. Also, add descriptive attribute if any on the relationship.

LEC-9: SQL in 1-Video

1. **SQL**: Structured Query Language, used to access and manipulate data.
2. SQL used **CRUD** operations to communicate with DB.
    1. **CREATE** - execute INSERT statements to insert new tuple into the relation.
    2. **READ** - Read data already in the relations.
    3. **UPDATE** - Modify already inserted data in the relation.
    4. **DELETE** - Delete specific data point/tuple/row or multiple rows.
3. **SQL is not DB, is a query language.**
4. What is **RDBMS**? (Relational Database Management System)
    1. Software that enable us to implement designed relational model.
    2. e.g., MySQL, MS SQL, Oracle, IBM etc.
    3. Table/Relation is the simplest form of data storage object in R-DB.
    4. **MySQL** is open-source RDBMS, and it uses SQL for all CRUD operations
5. **MySQL** used client-server model, where client is CLI or frontend that used services provided by MySQL server.
6. **Difference between SQL and MySQL**
    1. SQL is Structured Query language used to perform CRUD operations in R-DB, while MySQL is a RDBMS used to store, manage and administrate DB (provided by itself) using SQL.

SQL DATA TYPES (Ref: https://www.w3schools.com/sql/sql_datatypes.asp)

1. In SQL DB, data is stored in the form of tables.
2. Data can be of different types, like INT, CHAR etc.

DATATYPE	Description
CHAR	string(0-255), string with size = (0, 255], e.g., CHAR(251)
VARCHAR	string(0-255)
TINYTEXT	String(0-255)
TEXT	string(0-65535)
BLOB	string(0-65535)
MEDIUMTEXT	string(0-16777215)
MEDIUMBLOB	string(0-16777215)
LONGTEXT	string(0-4294967295)
LONGBLOB	string(0-4294967295)
TINYINT	integer(-128 to 127)
SMALLINT	integer(-32768 to 32767)
MEDIUMINT	integer(-8388608 to 8388607)
INT	integer(-2147483648 to 2147483647)
BIGINT	integer (-9223372036854775808 to 9223372036854775807)
FLOAT	Decimal with precision to 23 digits
DOUBLE	Decimal with 24 to 53 digits

DATATYPE	Description
DECIMAL	Double stored as string
DATE	YYYY-MM-DD
DATETIME	YYYY-MM-DD HH:MM:SS
TIMESTAMP	YYYYMMDDHHMMSS
TIME	HH:MM:SS
ENUM	One of the preset values
SET	One or many of the preset values
BOOLEAN	0/1
BIT	e.g., BIT(n), n upto 64, store values in bits.

3. Size: TINY < SMALL < MEDIUM < INT < BIGINT.
4. **Variable length Data types** e.g., VARCHAR, are better to use as they occupy space equal to the actual data size.
5. Values can also be unsigned e.g., INT UNSIGNED.
6. **Types of SQL commands:**
    1. **DDL** (data definition language): defining relation schema.
        1. **CREATE**: create table, DB, view.
        2. **ALTER TABLE**: modification in table structure. e.g, change column datatype or add/remove columns.
        3. **DROP**: delete table, DB, view.
        4. **TRUNCATE**: remove all the tuples from the table.
        5. **RENAME**: rename DB name, table name, column name etc.
    2. **DRL/DQL** (data retrieval language / data query language): retrieve data from the tables.
        1. **SELECT**
    3. **DML** (data modification language): use to perform modifications in the DB
        1. **INSERT**: insert data into a relation
        2. **UPDATE**: update relation data.
        3. **DELETE**: delete row(s) from the relation.
    4. **DCL** (Data Control language): grant or revoke authorities from user.
        1. **GRANT**: access privileges to the DB
        2. **REVOKE**: revoke user access privileges.
    5. **TCL** (Transaction control language): to manage transactions done in the DB
        1. **START TRANSACTION**: begin a transaction
        2. **COMMIT**: apply all the changes and end transaction
        3. **ROLLBACK**: discard changes and end transaction
        4. **SAVEPOINT**: checkout within the group of transactions in which to rollback.

MANAGING DB (DDL)

1. **Creation of DB**
    1. **CREATE DATABASE IF NOT EXISTS db-name;**
    2. **USE** db-name; //need to execute to choose on which DB CREATE TABLE etc commands will be executed. //make switching between DBs possible.
    3. **DROP** DATABASE IF EXISTS db-name; //dropping database.
    4. **SHOW** DATABASES; //list all the DBs in the server.
    5. **SHOW** TABLES; //list tables in the selected DB.

DATA RETRIEVAL LANGUAGE (DRL)

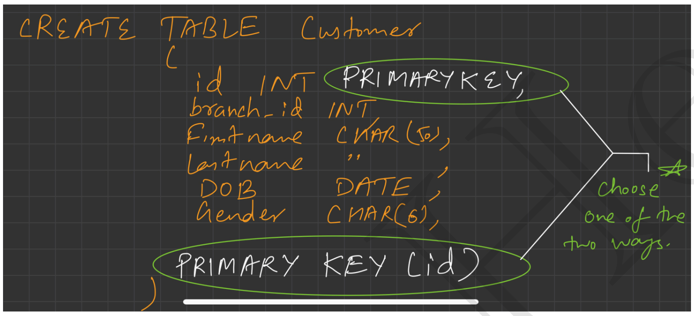
1. Syntax: SELECT <set of column names> FROM <table_name>;
2. Order of execution from RIGHT to LEFT.
3. Q. Can we use SELECT keyword without using FROM clause?
    1. Yes, using DUAL Tables.
    2. Dual tables are dummy tables created by MySQL, help users to do certain obvious actions without referring to user defined tables.
    3. e.g., SELECT 55 + 11;
       SELECT now();
       SELECT ucase(); etc.
4. **WHERE**
    1. Reduce rows based on given conditions.
    2. E.g., SELECT * FROM customer WHERE age > 18;
5. **BETWEEN**
    1. SELECT * FROM customer WHERE age between 0 **AND** 100;
    2. In the above e.g., 0 and 100 are inclusive.
6. **IN**
    1. Reduces **OR** conditions;
    2. e.g., SELECT * FROM officers WHERE officer_name IN ('Lakshay', 'Maharana Pratap', 'Deepika');
7. **AND/OR/NOT**
    1. AND: WHERE cond1 AND cond2
    2. OR: WHERE cond1 OR cond2
    3. NOT: WHERE col_name NOT IN (1,2,3,4);
8. **IS NULL**
    1. e.g., SELECT * FROM customer WHERE prime_status is NULL;
9. **Pattern Searching / Wildcard ('%', '_')**
    1. '%', any number of character from 0 to n. Similar to '*' asterisk in regex.
    2. '_', only one character.
    3. SELECT * FROM customer WHERE name **LIKE** '%p_';
10. **ORDER BY**
    1. Sorting the data retrieved using **WHERE** clause.
    2. ORDER BY <column-name> DESC;
    3. DESC = Descending and ASC = Ascending
    4. e.g., SELECT * FROM customer ORDER BY name DESC;
11. **GROUP BY**
    1. GROUP BY Clause is used to collect data from multiple records and group the result by one or more column. It is generally used in a SELECT statement.
    2. Groups into category based on column given.
    3. SELECT c1, c2, c3 FROM sample_table WHERE cond GROUP BY c1, c2, c3.
    4. All the column names mentioned after SELECT statement shall be repeated in GROUP BY, in order to successfully execute the query.
    5. Used with aggregation functions to perform various actions.
        1. COUNT()
        2. SUM()
        3. AVG()
        4. MIN()
        5. MAX()
12. **DISTINCT**
    1. Find distinct values in the table.
    2. SELECT DISTINCT(col_name) FROM table_name;
    3. GROUP BY can also be used for the same
        1. "Select col_name from table GROUP BY col_name;" same output as above DISTINCT query.

        2. SQL is smart enough to realise that if you are using GROUP BY and not using any aggregation function, then you mean "DISTINCT".

13. **GROUP BY HAVING**
    1. Out of the categories made by GROUP BY, we would like to know only particular thing (cond).
    2. Similar to WHERE.
    3. Select COUNT(cust_id), country from customer GROUP BY country HAVING COUNT(cust_id) > 50;
    4. WHERE vs HAVING
        1. Both have same function of filtering the row base on certain conditions.
        2. WHERE clause is used to filter the rows from the table based on specified condition
        3. HAVING clause is used to filter the rows from the groups based on the specified condition.
        4. HAVING is used after GROUP BY while WHERE is used before GROUP BY clause.
        5. If you are using HAVING, GROUP BY is necessary.
        6. WHERE can be used with SELECT, UPDATE & DELETE keywords while GROUP BY used with SELECT.

CONSTRAINTS (DDL)

1. **Primary Key**



    1. PK is not null, unique and only one per table.
2. **Foreign Key**
    1. FK refers to PK of other table.
    2. Each relation can having any number of FK.
    3. CREATE TABLE ORDER (
       id INT PRIMARY KEY,
       delivery_date DATE,
       order_placed_date DATE,
       cust_id INT,
       FOREIGN KEY (cust_id) REFERENCES customer(id)
       );
3. **UNIQUE**
    1. Unique, can be null, table can have multiple unique attributes.
    2. CREATE TABLE customer (
       …
       email VARCHAR(1024) UNIQUE,
       …
       );
4. **CHECK**
    1. CREATE TABLE customer (
       …
       CONSTRAINT age_check CHECK (age > 12),
       …
       );
    2. "age_check", can also avoid this, MySQL generates name of constraint automatically.

5. **DEFAULT**
    1. Set default value of the column.
    2. CREATE TABLE account (
       ...
       saving-rate DOUBLE NOT NULL DEFAULT 4.25,
       ...
       );
6. An attribute can be **PK and FK both** in a table.
7. **ALTER OPERATIONS**
    1. Changes schema
    2. **ADD**
        1. **Add new column.**
        2. ALTER TABLE table_name ADD new_col_name datatype ADD new_col_name_2 datatype;
        3. e.g., ALTER TABLE customer ADD age INT NOT NULL;
    3. **MODIFY**
        1. **Change datatype of an attribute.**
        2. ALTER TABLE table-name MODIFY col-name col-datatype;
        3. E.g., VARCHAR TO CHAR
           ALTER TABLE customer MODIFY name CHAR(1024);
    4. **CHANGE COLUMN**
        1. **Rename column name.**
        2. ALTER TABLE table-name CHANGE COLUMN old-col-name new-col-name new-col-datatype;
        3. e.g., ALTER TABLE customer CHANGE COLUMN name customer-name VARCHAR(1024);
    5. **DROP COLUMN**
        1. **Drop a column completely.**
        2. ALTER TABLE table-name DROP COLUMN col-name;
        3. e.g., ALTER TABLE customer DROP COLUMN middle-name;
    6. **RENAME**
        1. **Rename table name itself.**
        2. ALTER TABLE table-name RENAME TO new-table-name;
        3. e.g., ALTER TABLE customer RENAME TO customer-details;

DATA MANIPULATION LANGUAGE (DML)

1. **INSERT**
    1. INSERT INTO table-name(col1, col2, col3) VALUES (v1, v2, v3), (val1, val2, val3);
2. **UPDATE**
    1. UPDATE table-name SET col1 = 1, col2 = 'abc' WHERE id = 1;
    2. Update multiple rows e.g.,
        1. UPDATE student SET standard = standard + 1;
    3. **ON UPDATE CASCADE**
        1. Can be added to the table while creating constraints. Suppose there is a situation where we have two tables such that primary key of one table is the foreign key for another table. if we update the primary key of the first table then using the ON UPDATE CASCADE foreign key of the second table automatically get updated.
3. **DELETE**
    1. DELETE FROM table-name WHERE id = 1;
    2. DELETE FROM table-name; //all rows will be deleted.
    3. **DELETE CASCADE** - (**to overcome DELETE constraint of Referential constraints**)
        1. What would happen to child entry if parent table's entry is deleted?
        2. CREATE TABLE ORDER (
           order_id int PRIMARY KEY,
           delivery_date DATE,
           cust_id INT,

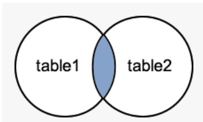
FOREIGN KEY(cust_id) REFERENCES customer(id) ON DELETE CASCADE
);

3. **ON DELETE NULL** - (**can FK have null values?**)
    1. CREATE TABLE ORDER (
       order_id int PRIMARY KEY,
       delivery_date DATE,
       cust_id INT,
       FOREIGN KEY(cust_id) REFERENCES customer(id) ON DELETE SET NULL
       );

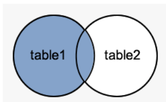
4. **REPLACE**
    1. Primarily used for already present tuple in a table.
    2. As UPDATE, using REPLACE with the help of WHERE clause in PK, then that row will be replaced.
    3. As INSERT, if there is no duplicate data new tuple will be inserted.
    4. REPLACE INTO student (id, class) VALUES(4, 3);
    5. REPLACE INTO table SET col1 = val1, col2 = val2;

JOINING TABLES

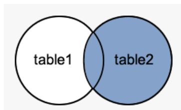
1. All **RDBMS** are relational in nature, we refer to other tables to get meaningful outcomes.
2. FK are used to do reference to other table.
3. **INNER JOIN**
    1. Returns a resultant table that has matching values from both the tables or all the tables.
    2. SELECT column-list FROM table1 INNER JOIN table2 ON condition1
       INNER JOIN table3 ON condition2
       ...;



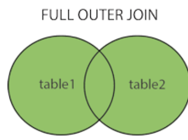
3. **Alias in MySQL (AS)**
    1. Aliases in MySQL is used to give a temporary name to a table or a column in a table for the purpose of a particular query. It works as a nickname for expressing the tables or column names. It makes the query short and neat.
    2. SELECT col_name AS alias_name FROM table_name;
    3. SELECT col_name1, col_name2,... FROM table_name AS alias_name;
4. **OUTER JOIN**
    1. **LEFT JOIN**
        1. This returns a resulting table that all the data from left table and the matched data from the right table.
        2. SELECT columns FROM table LEFT JOIN table2 ON Join_Condition;



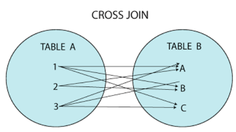
    2. **RIGHT JOIN**
        1. This returns a resulting table that all the data from right table and the matched data from the left table.
        2. SELECT columns FROM table RIGHT JOIN table2 ON join_cond;



    3. **FULL JOIN**
        1. This returns a resulting table that contains all data when there is a match on left or right table data.
        2. **Emulated** in MySQL using LEFT and RIGHT JOIN.
        3. LEFT JOIN UNION RIGHT JOIN.
        4. SELECT columns FROM table1 as t1 LEFT JOIN table2 as t2 ON t1.id = t2.id
           UNION
           SELECT columns FROM table1 as t1 RIGHT JOIN table2 as t2 ON t1.id = t2.id;
        5. UNION ALL, can also be used this will duplicate values as well while UNION gives unique values.



5. **CROSS JOIN**
    1. This returns all the cartesian products of the data present in both tables. Hence, all possible variations are reflected in the output.
    2. Used rarely in practical purpose.
    3. Table-1 has 10 rows and table-2 has 5, then resultant would have 50 rows.
    4. SELECT column-lists FROM table1 CROSS JOIN table2;



6. **SELF JOIN**

    1. It is used to get the output from a particular table when the same table is joined to itself.
    2. Used very less.
    3. Emulated using INNER JOIN.
    4. SELECT columns FROM table as t1 INNER JOIN table as t2 ON t1.id = t2.id;
7. **Join without using join keywords.**
    1. SELECT * FROM table1, table2 WHERE condition;
    2. e.g., SELECT artist_name, album_name, year_recordedFROM artist, albumWHERE artist.id = album.artist_id;

SET OPERATIONS

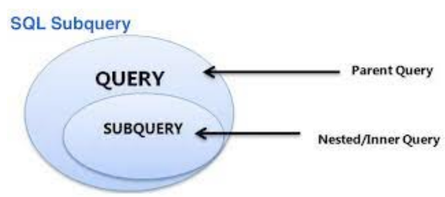
1. Used to combine multiple select statements.
2. Always gives distinct rows.

JOIN	SET Operations
Combines multiple tables based on matching condition.	Combination is resulting set from two or more SELECT statements.
Column wise combination.	Row wise combination.
Data types of two tables can be different.	Datatypes of corresponding columns from each table should be the same.
Can generate both distinct or duplicate rows.	Generate distinct rows.
The number of column(s) selected may or may not be the same from each table.	The number of column(s) selected must be the same from each table.
Combines results horizontally.	Combines results vertically.

3. **UNION**
    1. Combines two or more SELECT statements.
    2. SELECT * FROM table1
       UNION
       SELECT * FROM table2;
    3. Number of column, order of column must be same for table1 and table2.
4. **INTERSECT**
    1. Returns common values of the tables.
    2. Emulated.
    3. SELECT DISTINCT column-list FROM table-1 INNER JOIN table-2 USING(join_cond);
    4. SELECT DISTINCT * FROM table1 INNER JOIN table2 ON USING(id);
5. **MINUS**
    1. This operator returns the distinct row from the first table that does not occur in the second table.
    2. Emulated.
    3. SELECT column_list FROM table1 LEFT JOIN table2 ON condition WHERE table2.column_name IS NULL;
    4. e.g., SELECT id FROM table-1 LEFT JOIN table-2 USING(id) WHERE table-2.id IS NULL;

SUB QUERIES

1. Outer query depends on inner query.
2. Alternative to joins.
3. Nested queries.
4. SELECT column_list (s) FROM table_name WHERE column_name OPERATOR (SELECT column_list (s) FROM table_name [WHERE]);
5. e.g., SELECT * FROM table1 WHERE col1 IN (SELECT col1 FROM table1);
6. Sub queries exist mainly in 3 clauses
    1. Inside a WHERE clause.



2. Inside a FROM clause.
3. Inside a SELECT clause.

7. **Subquery using FROM clause**
    1. SELECT MAX(rating) FROM (SELECT * FROM movie WHERE country = 'India') as temp;
8. **Subquery using SELECT**
    1. SELECT (SELECT column_list(s) FROM T_name WHERE condition), columnList(s) FROM T2_name WHERE condition;
9. **Derived Subquery**
    1. SELECT columnLists(s) FROM (SELECT columnLists(s) FROM table_name WHERE [condition]) as new_table_name;
10. **Co-related sub-queries**
    1. With a normal nested subquery, the inner SELECT query runs first and executes once, returning values to be used by the main query. A correlated subquery, however, executes once for each candidate row considered by the outer query. In other words, the inner query is driven by the outer query.

```
SELECT column1, column2, ....
FROM table1 as outer
WHERE column1 operator
                    (SELECT column1, column2
                     FROM table2
                     WHERE expr1 =
                               outer.expr2);
```

JOIN VS SUB-QUERIES

JOINS	SUBQUERIES
Faster	Slower
Joins maximise calculation burden on DBMS	Keeps responsibility of calculation on user.
Complex, difficult to understand and implement	Comparatively easy to understand and implement.
Choosing optimal join for optimal use case is difficult	Easy.

MySQL VIEWS

1. A view is a database object that has no values. Its contents are based on the base table. It contains rows and columns similar to the real table.
2. In MySQL, the View is a **virtual table** created by a query by joining one or more tables. It is operated similarly to the base table but does not contain any data of its own.
3. The View and table have one main difference that the views are definitions built on top of other tables (or views). If any changes occur in the underlying table, the same changes reflected in the View also.
4. CREATE VIEW view_name AS SELECT columns FROM tables [WHERE conditions];
5. ALTER VIEW view_name AS SELECT columns FROM table WHERE conditions;
6. DROP VIEW IF EXISTS view_name;
7. CREATE VIEW Trainer AS SELECT c.course_name, c.trainer, t.email FROM courses c, contact t WHERE c.id = t.id; (View using Join clause).

NOTE: We can also import/export table schema from files (.csv or json).

LEC-11: Normalisation

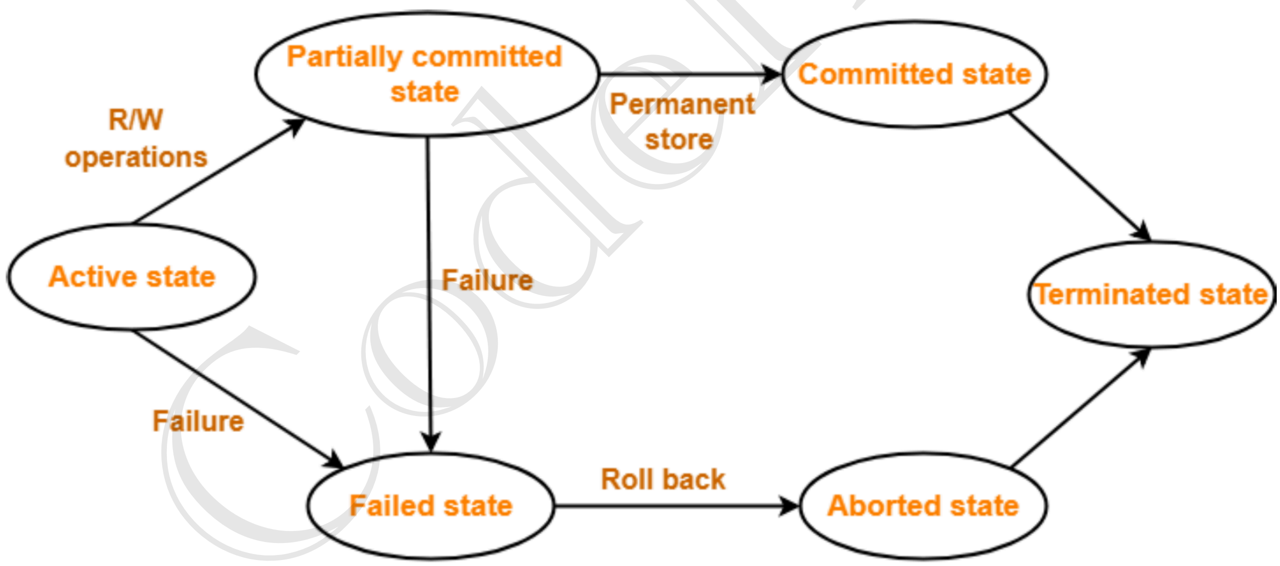
1. **Normalisation** is a step towards DB optimisation.
2. **Functional Dependency (FD)**
    1. It's a relationship between the primary key attribute (usually) of the relation to that of the other attribute of the relation.
    2. X -> Y, the left side of FD is known as a **Determinant**, the right side of the production is known as a **Dependent**.
    3. **Types of FD**
        1. **Trivial FD**
            1. A → B has trivial functional dependency if B is a subset of A. A->A, B->B are also Trivial FD.
        2. **Non-trivial FD**
            1. A → B has a non-trivial functional dependency if B is not a subset of A. [A intersection B is NULL].
    4. **Rules of FD (Armstrong's axioms)**
        1. **Reflexive**
            1. If 'A' is a set of attributes and 'B' is a subset of 'A'. Then, A→ B holds.
            2. If A ⊇ B then A → B.
        2. **Augmentation**
            1. If B can be determined from A, then adding an attribute to this functional dependency won't change anything.
            2. If A→ B holds, then AX→ BX holds too. 'X' being a set of attributes.
        3. **Transitivity**
            1. If A determines B and B determines C, we can say that A determines C.
            2. if A→ B and B→ C then A→ C.
3. **Why Normalisation?**
    1. To avoid redundancy in the DB, not to store redundant data.
4. **What happen if we have redundant data?**
    1. Insertion, deletion and updation anomalies arises.
5. **Anomalies**
    1. Anomalies means abnormalities, there are three types of anomalies introduced by data redundancy.
    2. **Insertion anomaly**
        1. When certain data (attribute) can not be inserted into the DB without the presence of other data.
    3. **Deletion anomaly**
        1. The delete anomaly refers to the situation where the deletion of data results in the unintended loss of some other important data.
    4. **Updation anomaly** (or modification anomaly)
        1. The update anomaly is when an update of a single data value requires multiple rows of data to be updated.
        2. Due to updation to many places, may be **Data inconsistency** arises, if one forgets to update the data at all the intended places.
    5. Due to these anomalies, **DB size increases** and **DB performance become very slow.**
    6. To rectify these anomalies and the effect of these of DB, we use **<u>Database optimisation technique</u>** called **NORMALISATION**.
6. **What is Normalisation?**
    1. Normalisation is used to minimise the redundancy from a relations. It is also used to eliminate undesirable characteristics like Insertion, Update, and Deletion Anomalies.
    2. Normalisation divides the composite attributes into individual attributes OR larger table into smaller and links them using relationships.
    3. The normal form is used to reduce redundancy from the database table.
7. **Types of Normal forms**
    1. **1NF**
        1. Every relation cell must have atomic value.
        2. Relation must not have multi-valued attributes.

2. **2NF**
    1. Relation must be in 1NF.
    2. There should not be any partial dependency.
        1. All non-prime attributes must be fully dependent on PK.
        2. Non prime attribute can not depend on the part of the PK.
3. **3NF**
    1. Relation must be in 2NF.
    2. No transitivity dependency exists.
        1. Non-prime attribute should not find a non-prime attribute.
4. **BCNF (Boyce-Codd normal form)**
    1. Relation must be in 3NF.
    2. FD: A -> B, A must be a super key.
        1. We must not derive prime attribute from any prime or non-prime attribute.

8. **Advantages of Normalisation**
    1. Normalisation helps to minimise data redundancy.
    2. Greater overall database organisation.
    3. Data consistency is maintained in DB.

LEC-12: Transaction

1. **Transaction**
    1. A unit of work done against the DB in a logical sequence.
    2. Sequence is very important in transaction.
    3. It is a logical unit of work that contains one or more SQL statements. The result of all these statements in a transaction either gets completed successfully (all the changes made to the database are permanent) or if at any point any failure happens it gets rollbacked (all the changes being done are undone.)
2. **ACID Properties**
    1. To ensure integrity of the data, we require that the DB system maintain the following properties of the transaction.
    2. **Atomicity**
        1. Either all operations of transaction are reflected properly in the DB, or none are.
    3. **Consistency**
        1. Integrity constraints must be maintained before and after transaction.
        2. DB must be consistent after transaction happens.
    4. **Isolation**
        1. Even though multiple transactions may execute concurrently, the system guarantees that, for every pair of transactions Ti and Tj, it appears to Ti that either Tj finished execution before Ti started, or Tj started execution after Ti finished. Thus, each transaction is unaware of other transactions executing concurrently in the system.
        2. Multiple transactions can happen in the system in isolation, without interfering each other.
    5. **Durability**
        1. After transaction completes successfully, the changes it has made to the database persist, even if there are system failures.
3. **Transaction states**



Transaction States in DBMS

    1. **Active state**
        1. The very first state of the life cycle of the transaction, all the read and write operations are being performed. If they execute without any error the T comes to Partially committed state. Although if any error occurs then it leads to a Failed state.
    2. **Partially committed state**
        1. After transaction is executed the changes are saved in the buffer in the main memory. If the changes made are permanent on the DB then the state will transfer to the committed state and if there is any failure, the T will go to Failed state.
    3. **Committed state**

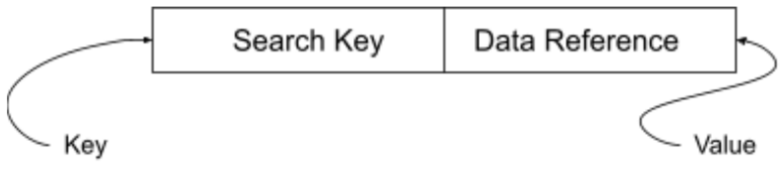
    1. When updates are made permanent on the DB. Then the T is said to be in the committed state. Rollback can't be done from the committed states. New consistent state is achieved at this stage.
4. **Failed state**
    1. When T is being executed and some failure occurs. Due to this it is impossible to continue the execution of the T.
5. **Aborted state**
    1. When T reaches the failed state, all the changes made in the buffer are reversed. After that the T rollback completely. T reaches abort state after rollback. DB's state prior to the T is achieved.
6. **Terminated state**
    1. A transaction is said to have terminated if has either committed or aborted.

LEC-13: How to implement Atomicity and Durability in Transactions

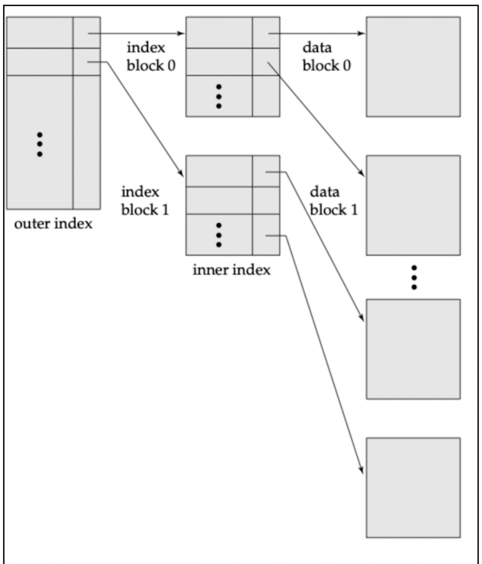
1. **Recovery Mechanism Component** of DBMS supports **atomicity and durability.**
2. **Shadow-copy scheme**
    1. Based on making copies of DB (aka, **shadow copies)**.
    2. Assumption only one Transaction (T) is active at a time.
    3. A pointer called **db-pointer** is maintained on the **disk**; which at any instant points to current copy of DB.
    4. T, that wants to update DB first creates a complete copy of DB.
    5. All further updates are done on new DB copy leaving the original copy (shadow copy) untouched.
    6. If at any point the **T has to be aborted** the system deletes the new copy. And the old copy is not affected.
    7. If T success, it is committed as,
        1. OS makes sure all the pages of the new copy of DB written on the disk.
        2. DB system updates the db-pointer to point to the new copy of DB.
        3. New copy is now the current copy of DB.
        4. The old copy is deleted.
        5. The T is said to have been **COMMITTED** at the point where the updated db-pointer is written to disk.
    8. **Atomicity**
        1. If T fails at any time before db-pointer is updated, the old content of DB are not affected.
        2. T abort can be done by just deleting the new copy of DB.
        3. Hence, either all updates are reflected or none.
    9. **Durability**
        1. Suppose, system fails are any time before the updated db-pointer is written to disk.
        2. When the system restarts, it will read db-pointer & will thus, see the original content of DB and none of the effects of T will be visible.
        3. T is assumed to be successful only when db-pointer is updated.
        4. If **system fails after** db-pointer has been updated. Before that all the pages of the new copy were written to disk. Hence, when system restarts, it will read new DB copy.
    10. The implementation is dependent on write to the db-pointer being atomic. Luckily, disk system provide atomic updates to entire block or at least a disk sector. So, we make sure db-pointer lies entirely in a single sector. By storing db-pointer at the beginning of a block.
    11. **Inefficient**, as entire DB is copied for every Transaction.
3. **Log-based recovery methods**
    1. The log is a sequence of records. Log of each transaction is maintained in some **stable storage** so that if any failure occurs, then it can be recovered from there.
    2. If any operation is performed on the database, then it will be recorded in the log.
    3. But the process of storing the logs should be done **before** the actual transaction is applied in the database.
    4. **Stable storage** is a classification of computer data storage technology that guarantees atomicity for any given write operation and allows software to be written that is robust against some hardware and power failures.
    5. **Deferred DB Modifications**
        1. Ensuring **atomicity** by recording all the DB modifications in the log but deferring the execution of all the write operations until the final action of the T has been executed.
        2. Log information is used to execute deferred writes when T is completed.
        3. If system crashed before the T completes, or if T is aborted, the information in the logs are ignored.
        4. If T completes, the records associated to it in the log file are used in executing the deferred writes.
        5. If failure occur while this updating is taking place, we preform redo.
    6. **Immediate DB Modifications**
        1. DB modifications to be output to the DB while the T is still in active state.
        2. DB modifications written by active T are called uncommitted modifications.
        3. In the event of crash or T failure, system uses old value field of the log records to restore modified values.
        4. Update takes place only after log records in a stable storage.
        5. Failure handling
            1. System failure before T completes, or if T aborted, then old value field is used to undo the T.
            2. If T completes and system crashes, then new value field is used to redo T having commit logs in the logs.

LEC-14: Indexing in DBMS

1. **Indexing** is used to **optimise the performance** of a database by minimising the number of disk accesses required when a query is processed.
2. The index is a type of **data structure.** It is used to locate and access the data in a database table quickly.
3. **Speeds up operation** with read operations like **SELECT** queries, **WHERE** clause etc.
4. **Search Key**: Contains copy of primary key or candidate key of the table or something else.
5. **Data Reference**: Pointer holding the address of disk block where the value of the corresponding key is stored.
6. Indexing is **optional**, but increases access speed. It is not the primary mean to access the tuple, it is the secondary mean.



7. **Index file is always sorted.**
8. **Indexing Methods**
    1. **Primary Index (Clustering Index)**
        1. A file may have several indices, on different search keys. If the data file containing the records is sequentially ordered, a Primary index is an index whose search key also defines the sequential order of the file.
        2. **NOTE**: The term primary index is sometimes used to mean an index on a primary key. However, such usage is **nonstandard** and **should be avoided.**
        3. All files are ordered sequentially on some search key. It could be Primary Key or non-primary key.
        4. **Dense And Sparse Indices**
            1. **Dense Index**
                1. The dense index contains an index record for every search key value in the data file.
                2. The index record contains the search-key value and a pointer to the first data record with that search-key value. The rest of the records with the same search-key value would be stored sequentially after the first record.
                3. It needs more space to store index record itself. The index records have the search key and a pointer to the actual record on the disk.
            2. **Sparse Index**
                1. An index record appears for only some of the search-key values.
                2. Sparse Index helps you to resolve the issues of dense Indexing in DBMS. In this method of indexing technique, a range of index columns stores the same data block address, and when data needs to be retrieved, the block address will be fetched.
        5. Primary Indexing can be based on Data file is sorted w.r.t Primary Key attribute or non-key attributes.
        6. **Based on Key attribute**
            1. Data file is sorted w.r.t primary key attribute.
            2. PK will be used as search-key in Index.
            3. Sparse Index will be formed i.e., no. of entries in the index file = no. of blocks in datafile.
        7. **Based on Non-Key attribute**
            1. Data file is sorted w.r.t non-key attribute.
            2. No. Of entries in the index = unique non-key attribute value in the data file.
            3. This is dense index as, all the unique values have an entry in the index file.
            4. E.g., Let's assume that a company recruited many employees in various departments. In this case, clustering indexing in DBMS should be created for all employees who belong to the same dept.
        8. **Multi-level Index**
            1. Index with two or more levels.
            2. If the single level index become enough large that the binary search it self would take much time, we can break down indexing into multiple levels.



Two-level sparse index.

    2. **Secondary Index (Non-Clustering Index)**
        1. Datafile is unsorted. Hence, Primary Indexing is not possible.
        2. Can be done on key or non-key attribute.
        3. Called secondary indexing because normally one indexing is already applied.
        4. No. Of entries in the index file = no. of records in the data file.
        5. It's an example of Dense index.

9. **Advantages of Indexing**
   1. Faster access and retrieval of data.
   2. IO is less.
10. **Limitations of Indexing**
   1. Additional space to store index table
   2. Indexing Decrease performance in INSERT, DELETE, and UPDATE query.