# Basic Database Concepts

|  |  |
| --- | --- |
| Data | Raw, unorganized, unprocessed facts. E.g. 5 |
| Information | Processed, organized, and structured data. E.g. John’s age is 5 |
| Information System | Integrated set of components for collecting, storing, and processing data to convert data into information |
| Database | Organized data store. Can be a component in an information system |
| State | * Complete DB with data * Entering, modifying, or deleting information changes the database state * When a transaction is not fully executed, changes are rolled back to the previous state |
| Operation | A simple query consisting of minimal functions on a database |
| Transaction | * Complex query with multiple operations * Logical unit of DB operation performed by DBMS * Independent of other transactions * A set of operations that must all be completed, or none processed at all * ACID Properties   + Atomic (Indivisible unit; Performed in entirety or not at all)   + Consistent (Executed from beginning to end without interference; Take DB from one consistent state to another)   + Isolated (Executed independent of other transactions)   + Durable (Change applied by a committed transaction persists in DB) |

## Need for Databases

* Efficient data management
* Perform multiple tasks easily
* Share data based on preferential access
* Stores, organizes, manages large information

## Data Consistency and Integrity

* Ensures that data cannot be written that would violate the database’s own rules for valid data
* Method: A**C**ID Transaction

## Concurrency

* Allows simultaneous access to database by multiple users
* Ability of DB to allow multiple users to affect multiple transactions
* Unlike spreadsheets where once a user opens, locked to other users

## Data Validation (Format Checking)

* Ensures databases are consistent, functional, and valuable
* Example
  + Enforce unique primary keys such as usernames
  + Hashed passwords have certain length
* Important to ensure data integrity

## Data Verification

* Check for inaccuracy and inconsistency
* Checks that data within DB entered/imported correctly from source
* E.g. Double entry of passwords

# RDBMS

|  |  |
| --- | --- |
| DBMS | * Software package to manipulate (create, read, update, and delete) data in DB * Provide interface between users and DB * Provides specific set of data based on user access privileges |
| RDBMS | * DB engine based on relational model * Tables may be related to other tables * Possible to run queries across tables |
| Schema | * Plans/Logical structure for DB * Defines how the data is organized and how the relations are associated * 3 Levels (Below) |

## Function and Tools of Databases

* Query, Update, Remove Data
* Data security and privacy system
  + Ensures transactions are authorized, regulated, and tracked
* Data Backup
  + In the case of failure
* Transaction Management (ACID)

## DBMS for Data Security

* Ensure integrity of data; Protected from deletion and corruption
  + Data validation
  + Data locking
* Authentication (UserID and Password)
* Authorization (Access Levels)
* Network monitoring & limiting volume of data accessed by a user

## Levels of Schema

### Conceptual

* High level data model
* Provides understanding of DB structure, semantics, interrelationships, and constraints
* Diagrammatic and simple representation in which non-specialized users can understand

### Logical

* Describes data in as much detail as possible without regard to how they will be physically implemented in DBMS
* Includes entities, attributes, keys
* Normalization included
* System independent mapping and tailoring schemas to specific DBMS
* Result is Data Definition Language (DDL) statements in language chosen by DBMS

### Physical

* Design specifications of stored DB in terms of physical file storage structures, record placements and indexes
* Translate entities into physical tables
* Attributes to use for columns
* Choosing most appropriate structure for DB files

## Data Dictionary

* Contains metadata which describes data structure, constraints, applications, authorization, users
* Stores
  + Description of schema
  + Physical DB design
  + Types of DB users with access rights
  + Changes made to the DB (Data provenance)
  + Relationship between DB transactions

## Data Definition Language

* Used to define schemas and create and modify the structure of database objects in a database
* Able to create, modify and delete tables, schema, views, and fields.

## Data Modelling

* Translating business logic into tables, columns, and relations
  + If entities, keys, and relationships not identified properly, cannot fulfil purpose
* Maintainability of DB
  + Normalization to reduce data duplication

|  |  |
| --- | --- |
| Table | * Set of data elements (values) * Model of columns (identifiable by name) and rows * Specified columns, unlimited rows |
| Record | * A row or tuple in a table * Composed of fields * Set of records constitutes file |
| Field | * Space allocated for a particular item of information * Smallest unit of information user can access * Is an attribute/column |
| Primary Key | * Uniquely identifies each record in a table * Not null |
| Secondary Key | * Candidate keys not chosen as primary keys |
| Foreign Key | * Uniquely identifies a row of another table * Defined in a 2nd table, refers to primary key in 1st table |
| Candidate Key | * A column or combination of columns that qualify as a primary key |
| Composite Primary Key | * Multiple attributes that together uniquely identify a record |
| (Inner) Join | * Establish connection between >= 2 DB tables based on matching columns |

## Entity Relationship Diagram

## Data Redundancy

* Same data found in multiple locations
* Takes up unnecessary space
* Duplication of effort: Single logical update multiple times
* Inconsistent files as updates may not be applied to all files 🡪 Inaccurate results

## Referential Integrity

* Ensures relations between tables remain consistent
* Refers to accuracy & consistency of data within a r/s
* FK must reference PK from parent table
* Links must be valid; FK must have actual value of key in parent table

## Normalization

### 1NF

* Atomic values in each cell
  + Each cell has only one value
* All columns and rows are unique

### 2NF

* Full functional dependency; Should not have partial dependency
* Attributes must be dependent on primary key
* Attributes must be entirely dependent on all parts of composite key, not just one column in the composite key

### 3NF

* No transitive dependencies
* Transitive dependency: When modifying non-key column affects other non-key column
* E.g. Full name and surname

## Usage of Views

* View: Virtual data derived but not explicitly stored in DB
* Different users require different perspectives/views of DB
* View is a subset of the database and may contain derived values
* E.g. School test grading system
  + Students can view only their own scores
  + Teachers can view all their student’s scores
  + Principal can view all student’s scores with summaries (e.g. average score)

## SQL

**SELECT <columns> FROM <table> WHERE <conditions>**

<columns>: Can be **\*** to represent ALL

Operators in <conditions>:

* =
* <> (Not equal to)
* >, < , >=, <=
* BETWEEN \_\_\_\_ AND \_\_\_\_\_

Other functions:

* SELECT DISTINCT \* FROM <table>
* SELECT MIN(<column>) FROM <table>
* SELECT MAX(<column>) FROM <table>
* SELECT AVG(<column>) FROM <table>
* SELECT SUM(<column>) FROM <table>

# Further DB Management

## Role of Database Administrator (DBA)

* Monitoring user access and security
* Installing, configuring, testing, and updating the DBMS
* Maintaining data standards and ensuring adherence to Data Protection Act
* Performance monitoring

## Database Recovery

* Log-based recovery
  + It is important that the logs are written prior to actual modification and stored on a stable storage media, which is failsafe
  + This is done by keeping the log file on stable storage media or when a transaction enters the system and starts execution, it writes a log about it;
* Back up files of the current database;
* Refer to them at any time to reconstitute the database

## Integrated DB Management Systems

* Relational DBMS
* Network model
* A collection of alphanumerical data and picture data
* Organized logically as a set of relations

## Personal Data Privacy Methods

* Encryption
* Ensure only a few trusted people have access to the data
* Monitor user access to data
* Do not collect unnecessary personal information

## Data Mining vs Data Matching

### Data Mining

* Set of automated techniques used to extract buried or previously unknown pieces of information from large databases
* Used in different contexts to achieve different goals
* Used in organisations to enable them to design effective sales campaigns, precise marketing plans and develop products to increase sales

### Data Matching

* Large scale comparison of records of files collected or held for different purposes with a view to identify matters of interest
* Economically feasible
* Conducted regularly particularly by government agencies
* Purpose: Detecting errors and illegal behaviour, locating individuals, facilitate debt collection