Database

Basics

1. **Importance of Data**
    1. ***"Data is the new oil"***
    2. *a database is a shared collection of logically related data ,description of these data ,designed to meet the information need of an organization*
    3. *data storage*
    4. *data analysis*
    5. *record keeping*
    6. *web applications*
2. *CRUD -- Create ,Retrieve, Update ,Delete*
3. **Properties of an ideal database**
    1. *Integrity -- Accuracy + Consistency*
    2. *Availability*
    3. *Security*
    4. *Independent of Application*
    5. *Concurrency*
4. **Types of Database**
    1. **Relational Databases** : *also known as SQL databases, these databases use a relational model to organize the data into tables with rows and columns eg **MySQL,Oracle,Postgres,SQL Server** basically all of the relational databases are row based databases. **Other names OLTP(Online Transactional Processing), Row based databases** used **for running web applications***.
    how row databases work?

For instance, let's take this Facebook_Friends data:

Facebook_Friends

Name	City	Age
Matt	Los Angeles	27
Dave	San Francisco	30
Tim	Oakland	33

This data would be stored on a disk in a row oriented database in order row by row like this:

Matt	Los Angeles	27	Dave	San Francisco	30	Tim	Oakland	33

Writing to Row Store Databases

Let's use the data stored in a database:

Matt	Los Angeles	27	Dave	San Francisco	30	Tim	Oakland	33

If we want to add a new record:

Jen	Vancouver	30

We can just append it to the end of the current data:

Matt	Los Angeles	27	Dave	San Francisco	30	Tim	Oakland	33	Jen	Vancouver	30

Row oriented databases are fast at retrieving a row or a set of rows but when performing an aggregation it brings extra data (columns) into memory which is slower than only selecting the columns that you are performing the aggregation on. In addition the number of disks the row oriented database might need to access is usually larger.

Extra data into Memory

Say we want to get the sum of ages from the Facebook_Friends data. To do this we will need to load all nine of these pieces of data into memory to then pull out the relevant data to do the aggregation.

Matt	Los Angeles	27	Dave	San Francisco	30	Tim	Oakland	33

This is wasted computing time.

disadvantages of row databases.

Number of Disks accessed

Let's assume a Disk can only hold enough bytes of data for three columns to be stored on each disk. In a row oriented database the table above would be stored as:

Disk 1		
Name	City	Age
Matt	Los Angeles	27

Disk 2		
Name	City	Age
Dave	San Francisco	30

Disk 3		
Name	City	Age
Tim	Oakland	33

To get the sum of all the people's ages the computer would need to look through all three disks and across all three columns in each disk in order to make this query.

2. **NoSQL Databases : ***these databases are designed to handle large amounts of unstructured data or semi structured data such as documents ,images, or videos eg** MongDB**
3. **Column Databases** : *these databases stores data in columns rather than rows making them well suited for analytical purposes and data warehousing. eg **Amazon Redshift , Google BigQuery**, **other names for these databases are OLAP,Datawarehouse, Snowflake** , used for **analysis purposes**.*
   how column databases work?

Facebook_Friends

Name	City	Age
Matt	Los Angeles	27
Dave	San Francisco	30
Tim	Oakland	33

A table is stored one column at a time in order row by row:

Matt	Dave	Tim	Los Angeles	San Francisco	Oakland	27	30	33

Writing to a Column Store Databases

If we want to add a new record:

Jen	Vancouver	30

We have to navigate around the data to plug each column in to where it should be.

Matt	Dave	Tim	Jen	Los Angeles	San Francisco	Oakland	Vancouver	27	30	33	30

Memory optimization in case of column databases

If the data was stored on a single disk it would have the same extra memory problem as a row oriented database, since it would need to bring everything into memory. However, column oriented databases will have significant benefits when stored on separate disks.

If we placed the table above into the similarly restricted three columns of data disk they would be stored like this:

Disk 1		
Name		
Matt	Dave	Tim

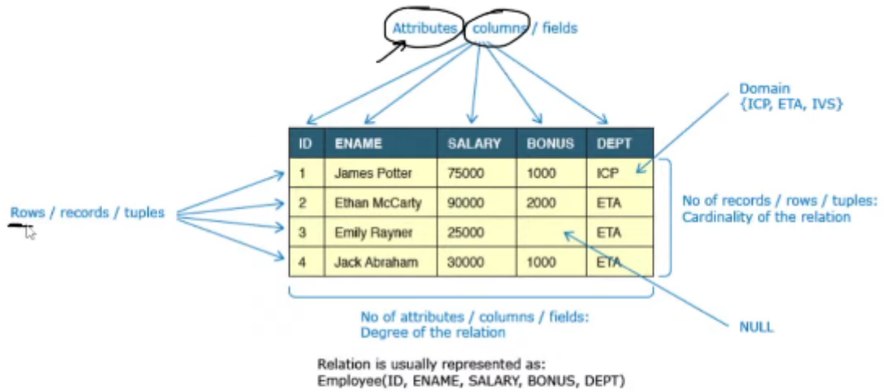
Disk 2		
City		
Los Angeles	San Francisco	Oakland

Disk 3		
Age		
27	30	33

4. **Graph Databases** : *these databases are used to query to graphs , and used in social media platforms. eg **Amazon Neptune, Neo4j***
5. **Key-Value Based Databases**: *generally we want aggregated function which are pre-calculated and can be shown instantly, this information is cached and can be shown instantly **eg Redis***

5. for further https://dataschool.com/data-modeling-101/row-vs-column-oriented-databases/
6. Relational Databases

1. Also known as SQL Databases, these databases use a relational model to organize data into tables with rows and columns.



    1. **rows** are called records or tuples
    2. **columns** are called attributes or fields
    3. **table** is called relation
7. What is DBMS?

    a database management system is a software system that provides the interfaces and tools needed to store , organize and manage data in databases. A DBMS acts as an intermediary between the application or user that access the data stored in databases
    Users
    |
    |
    application
    |
    |
    DBMS
    |
    |
    Operating System
8. Functions of DBMS
    1. Data Management
    2. Integrity -- maintain accuracy of data
    3. Concurrency -- simultaneous access of multiple users.
    4. Transaction -- modification to database must either be successful or must not happen at all.
    5. Security
    6. Utilities
9. Types of Keys

10. **Super key** : is a combination of columns that uniquely identifies the whole table within a RDBMS table.

Roll no	Name	Branch	email - id
01	Swayam	cse	swayam@gmail.com
02	rahul	ece	rahul@gmail.com
03	akhilesh	mec	akhilesh@gmail.com

in this table the combination of the columns can be used to uniquely identify the relation

1. roll no
2. email
3. roll no + name
4. roll no + email
5. name + branch + email
6. roll no + name + branch
7. roll no + name + branch + email

11. **Candidate Key**: a subset of of super key , it has no redundant attributes or columns that is it is at the simplest and can uniquely identify the relation
    1. roll no
    2. email

    these columns can be used to identify the whole relation and is non - redundant.
12. **Primary Key** : is a unique identifier for each tuple in a table . There can only be one primary key in a table and it cannot contain null values.

    `in a way we can say that super key is like the whole population who can stand in the election and candidate key are the candidates who stood in the election and primary key is the winner of the election`

    Conditions for primary key:

    1. null
    2. no duplicates

    good to have primary key

    1. numerical
    2. small
    3. consistent over time

       `roll no can be the primary key`

13. **Composite key** : is made up of combination of primary key that is made up of two or more attributes . it is made when a single key is not sufficient to uniquely identify a tuple in a table
14. **Surrogate Key** :

    in this type of table we cannot have a primary key through combination of any of these tuples so in order to form a primary key we introduce our own key which acts as a primary key and is called a **Surrogate Key**.

name	branch	cgpa
swayam	cse	9.8
akhilesh	mec	9.7
rahul	ece	9.8

sid here is introduced by us is acting as a primary key

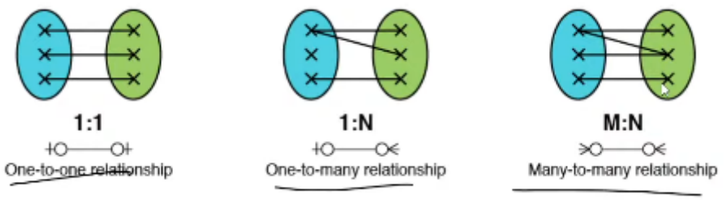
sid
1
2
3

name	branch	cgpa
swayam	cse	9.8
akhilesh	mec	9.7
rahul	ece	9.8

15. **Foreign Key** : a primary key from one table is used as primary key in another table , this is used to create relationship between tables.

16. **Cardinality of Relationships**

    in a database relationships refers to the number of occurrences of an entity in a relationship with another entity . Cardinality defines the number of instances of one entity that can be associated with a single instance of the related entity.



17. Drawbacks of Databases
    1. Complexity
    2. Cost
    3. Scalability
    4. Security
    5. Data Migration
    6. Flexibility