

## Define integrity constraint.

Integrity constraints are a set of rules that are used to maintain the quality of information.

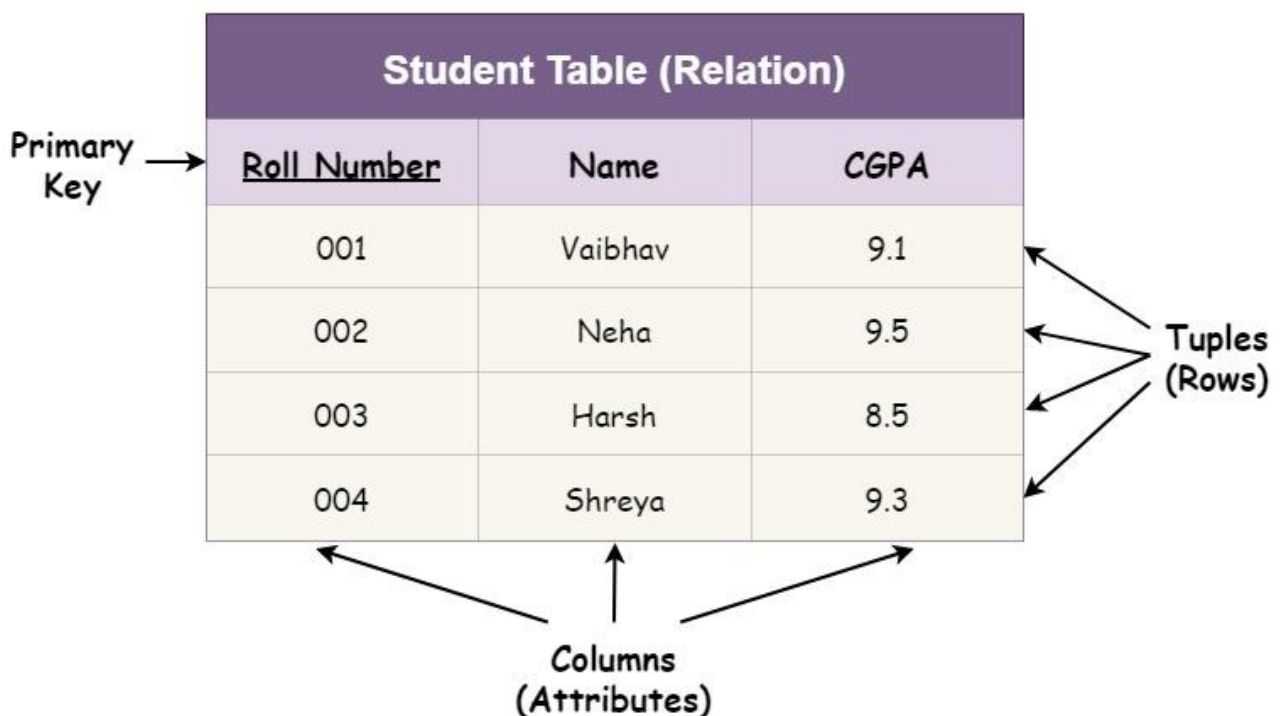
## Discuss relational model with the help of example.

The relational model represents how data is stored in relational databases.

Key features:

- **Attribute:** Attributes are the properties that define an entity.
- **Relation Schema:** It represents the name of the relation with its attributes.
- **Tuple:** Row in a relation is called a tuple.
- **Relation Instance:** The set of tuples of a relation at a particular instance of time.
- **Degree:** The number of attributes in the relation.
- **Cardinality:** The number of tuples in a relation.
- **Column:** Set of values for a particular attribute.
- **NULL Values:** Unknown value is called a NULL value.
- **Relation Key:** They are used to uniquely identify rows in a relation.

## Relational Model in DBMS



## Define data model?

Data Model refers to tools which are used to summarize the description of the database.

### Types of Relational Models

1. **Conceptual Data Model:** It is used in the requirement-gathering process

2. **Representational Data Model:** It is used to define the data and the relationships between them.
3. **Physical Data Model:** It is used to represent only the logical part of the database.

### Define normalization.

**Normalization** is the process of minimizing **redundancy** from a relation or set of relations.

### Explain query optimization?

Query Optimization is a technique of analysing and deciding an execution plan that computes the result of the query using less resources.

Two main objectives of Query Optimization are:

- Determine the optimal plan to access the database.
- Reduce the time required to execute the query.

There are following two methods of Query Optimization:

- **Cost Based Query Optimization:** In Cost Based Query Optimization, optimizer associates a numerical value (known as cost) for each step of feasible plan for a given query. Then, all these values are collectively analysed to get a cost estimate for that plan. After evaluating cost of all feasible plans, optimizer finds the plan with lowest cost estimate.
- **Adaptive Query Optimization:** In Adaptive Query Optimization, optimizer is allowed to make run time changes to the execution plans and can find new information to improve the optimizations. It is helpful when existing statistics are not sufficient to generate the plans.

### Explain semantic data model with the help of example.

It is a data model defined on a higher level that captures the database's semantic description, structure, and form. SDMs present a business-user-friendly perspective of the data.

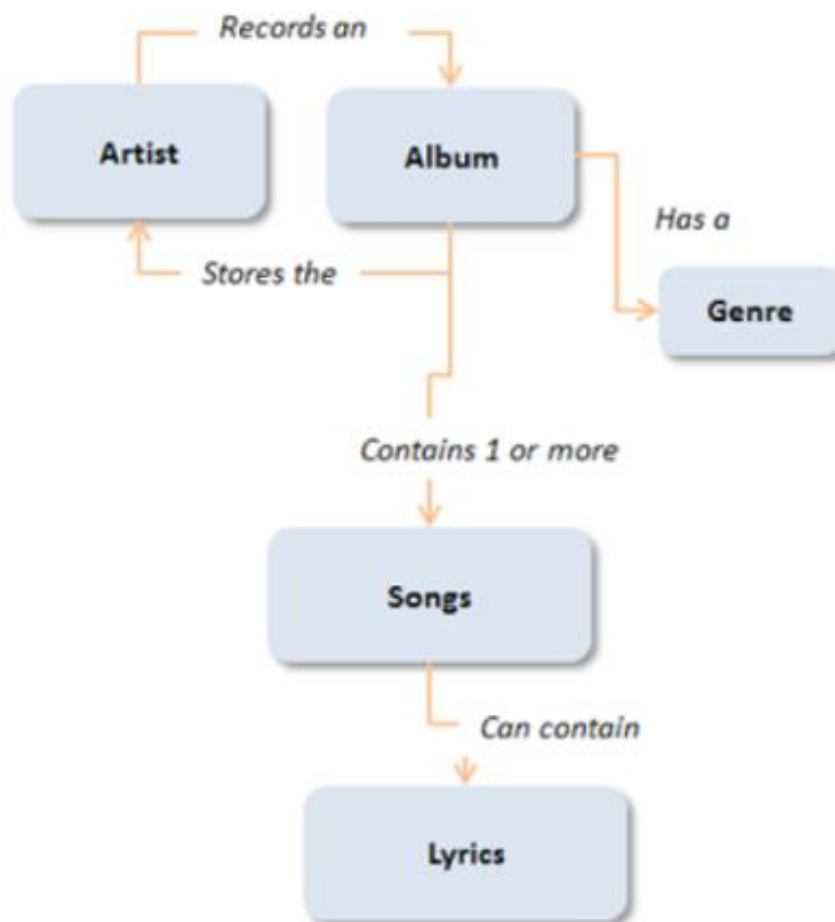
#### Advantages of SDM:

- Reveals relationships between instances
- Supports data visualization
- Supports application development
- Does not require technical knowledge about data models

#### Disadvantages of SDM:

- Uses graph data modeling technique
- Requires more practice and experience to create them

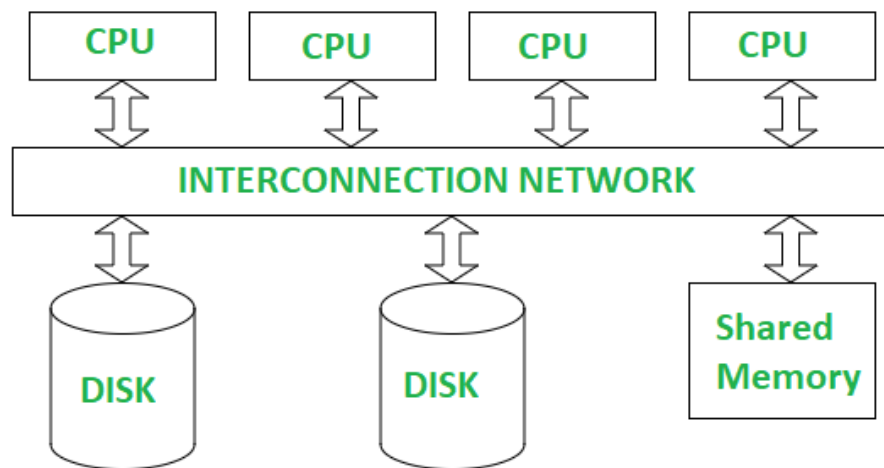
Example 1: This example visualizes the relationship between real-world objects in the music industry. Between each object are defined relationships and the direction of object dependence.



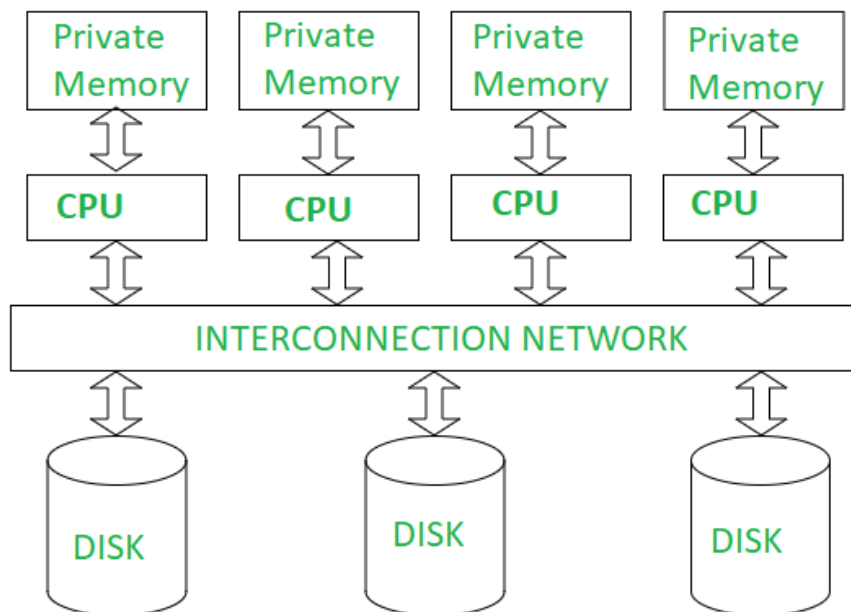
Discuss the design of parallel database.

There are three architectural designs for parallel database:

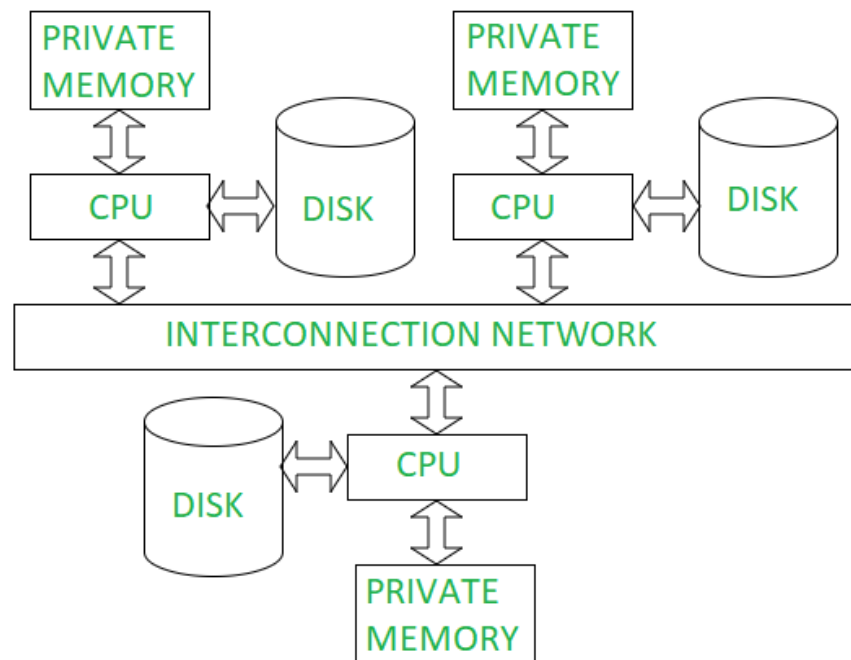
- 1. Shared Memory Architecture-** In Shared Memory Architecture, multiple CPUs are attached to an interconnection network. They share a single main memory and common disk arrays. A single copy of a multi-threaded OS and multi-threaded DBMS can support these multiple CPUs. It is also known as **Symmetric multiprocessing (SMP)**. This architecture is used in personal workstations.



**2. Shared Disk Architectures:** In Shared Disk Architecture, multiple CPUs are attached to an interconnection network. Each CPU has its own memory and all of them share the same disk. Each node has its own copy of the OS and DBMS. It is a loosely coupled architecture optimized for centralized applications. They are also known as **clusters**.



**3. Shared Nothing Architecture:** In this, multiple CPUs are attached to an interconnection network through a node. It is multiple processor architecture in which each processor has its own memory and disk storage. No two CPUs can access the same disk area. Memory or disk resources is not shared. It is also known as **Massively parallel processing (MPP)**.



**4. Hierarchical Architecture:** It is a combination of shared disk, shared memory and shared nothing architectures. It is scalable due to availability of more memory and many processors. But is costly to other architecture.

Explain the various algorithms for rule discovery.

- **Apriori Algorithm:** It is based on the concept of generating frequent itemsets and using them to derive association rules. Apriori works by iteratively generating candidate itemsets of increasing size and then pruning those that do not meet the minimum support threshold. This process continues until no more frequent itemsets can be generated. The algorithm derives association rules from the frequent itemsets, using metrics such as support and confidence to evaluate the strength of the rules.
- **Sampling Algorithm:** These algorithms work by randomly sampling a portion of the dataset and then applying traditional association rule mining techniques, such as the Apriori algorithm, to the sampled data. By analyzing a smaller subset of the data, sampling algorithms can be more efficient and scalable, especially for datasets that are too large to process in their entirety.
- **FP-Growth Algorithm:** The FP-Growth (Frequent Pattern Growth) algorithm is an alternative to the Apriori algorithm for mining frequent itemsets and association rules. FP-Growth works by constructing a compact data structure called the FP-tree, which represents the frequent itemsets in the dataset. This structure allows for efficient mining of frequent itemsets without the need for generating candidate itemsets. By

eliminating the need for candidate generation and pruning, FP-Growth can be more efficient than the Apriori algorithm, especially for datasets with a large number of transactions and items.

- **Partition Algorithm:** Partition-based algorithms are a class of algorithms that partition the dataset into smaller subsets or partitions and then mine association rules from each partition independently. These algorithms work by dividing the dataset into partitions based on some criteria, such as transaction IDs or item attributes. Association rules are then mined separately from each partition. Once association rules are derived from each partition, they can be combined or aggregated to produce the final set of rules for the entire dataset.