UNIT 4

TRANSACTION PROCESSING

Data Mining Tasks, OLAP and Multidimensional data analysis, Basic concept of Association Analysis and Cluster Analysis - Transaction processing v/s Analytic Processing- OLTP v/s OLAP- OLAP Operations - Data models for OLTP (ER model) and OLAP (Star & Snowflake Schema)

DATA MINING TASKS

Data mining tasks are majorly categorized into two categories: descriptive and predictive.

1. **Descriptive data mining**:
   Descriptive data mining offers a detailed description of the data, for example- it gives insight into what's going on inside the data without any prior idea. This demonstrates the common characteristics in the results. It includes any information to grasp what's going on in the data without a prior idea.
2. **Predictive Data Mining**:
   This allows users to consider features that are not specifically available. For example, the projection of the market analysis in the next quarters with the output of the previous quarters, In general, the predictive analysis forecasts or infers the features of the data previously available. For an instance: judging by the outcomes of medical records of a patient who suffers from some real illness.

Key Data Mining Tasks

1) Characterization and Discrimination

- **Data Characterization**: The characterization of data is a description of the general characteristics of objects in a target class which creates what are called characteristic rules.
  A database query usually computes the data applicable to a user-specified class and runs through a description component to retrieve the meaning of the data at various abstraction levels.
  Eg;-Bar maps, curves, and pie charts.
- **Data Discrimination**: Data discrimination creates a series of rules called discriminate rules that is simply a distinction between the two classes aligned with the goal class and the opposite class of the general characteristics of objects.

2) Prediction

To detect the inaccessible data, it uses regression analysis and detects the missing numeric values in the data. If the classmark is absent, so classification is used to render the prediction. Due to its relevance in business intelligence, the prediction is common. If the classmark is absent, so the prediction is performed using classification. There are two methods of predicting data. Due to its relevance in business intelligence, a prediction is common. The prediction of the classmark using the previously developed class model and the prediction of incomplete or incomplete data using prediction analysis are two ways of predicting data.

3) Classification

Classification is used to create data structures of predefined classes, as the model is used to classify new instances whose classification is not understood. The instances used to produce the model are known as data from preparation. A decision tree or set of classification rules is based on such a form of classification process that can be collected to identify future details, for example by classifying the possible compensation of the employee based on the classification of salaries of related employees in the company.

4) Association Analysis

The link between the data and the rules that bind them is discovered. And two or more data attributes are associated. It associates qualities that are transacted together regularly. They work out what are called the rules of partnerships that are commonly used in the study of stock baskets. To link the attributes, there are two elements. One is the trust that suggests the possibility of both associated together, and another helps, which informs of associations' past occurrence.

5) Outlier Analysis

Data components that cannot be clustered into a given class or cluster are outliers. They are often referred to as anomalies or surprises and are also very important to remember. Although in some contexts, outliers can be called noise and discarded, they can disclose useful information in other areas, and hence can be very important and beneficial for their study.

6) Cluster Analysis

Clustering is the arrangement of data in groups. Unlike classification, however, class labels are undefined in clustering and it is up to the clustering algorithm to find suitable classes. Clustering is often called unsupervised classification since provided class labels do not execute the classification. Many clustering methods are all based on the concept of maximizing the similarity (intra-class similarity) between objects of the same class and decreasing the similarity between objects in different classes (inter-class similarity).

7) Evolution & Deviation Analysis

We may uncover patterns and shifts in actions over time, with such distinct analysis, we can find features such as time-series results, periodicity, and similarities in patterns. Many technologies from space science to retail marketing can be found holistically in data processing and features.

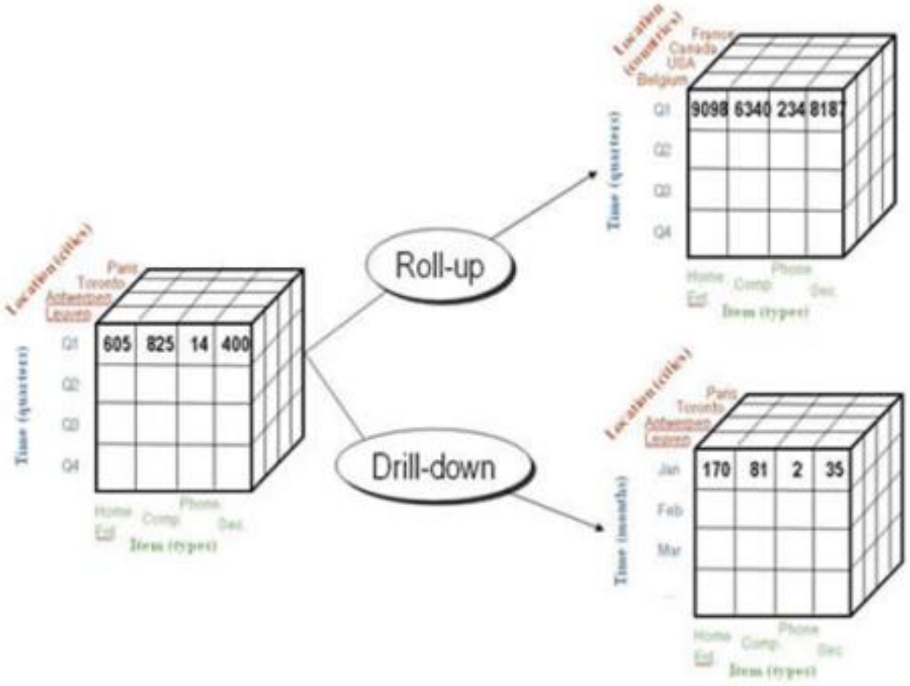
OLAP AND MULTIDIMENSIONAL DATA ANALYSIS

Most times used interchangeably, the terms Online Analytical Processing (OLAP) and data warehousing apply to decision support and business intelligence systems. OLAP systems help data warehouses to analyze the data effectively. The dimensional modeling in data warehousing primarily supports OLAP, which encompasses a greater category of business intelligence like relational database, data mining and report writing.

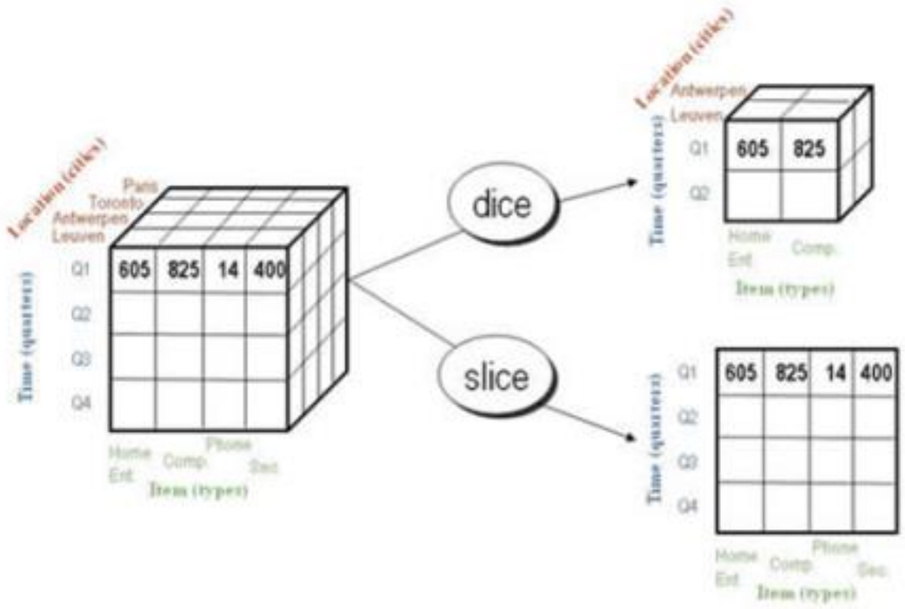
Many of the OLAP applications include sales reporting, marketing, business process management (BPM), forecasting, budgeting , creating finance reports and others. Each OLAP cube is presented through measures and dimensions. Measures refers to the numeric value categorized by dimensions. In below diagrams, dimensions are time, item type and courtiers/cities and the values inside them (605, 825, 14, 400) are measures.
The OLAP approach is used to analyze multidimensional data from multiple sources and perspectives. The three basic operations in OLAP are:

- **Roll-up (Consolidation)**
- **Drill-down**
- **Slicing and dicing**

Roll-up or consolidation refers to data aggregation and computation in one or more dimensions. It is actually performed on an OLAP cube. For instance, the cube with cities is rolled up to countries to depict the data with respect to time (in quarters) and item (type).



On the contrary, Drill-down operation helps users navigate through the data details. In the above example, drilling down enables users to analyze data in the three months of the first quarter separately. The data is divided with respect to cities, months (time) and item (type).



Slicing is an OLAP feature that allows taking out a portion of the OLAP cube to view specific data. For instance, in the above diagram, the cube is sliced to a two dimensional view showing Item(types) with respect to Quadrant (time). The location dimension is skipped here. In dicing, users can analyze data from different viewpoints. In the above diagram, the users create a sub cube and chose to view data for two Item types and two locations in two quadrants.

Multidimensional model (MOALP)

The databases that are configured for OLAP use multidimensional data model, enabling complex analysis and ad hoc queries at a rapid rate. The multidimensional data model is analogous to relational database model with a variation of having multidimensional structures for data organization and expressing relationships between the data. The data is stored in the form of cubes and can be accessed within the confines of each cube. Mostly, data warehousing supports two or three-dimensional cubes; however, there are more than three data dimensions depicted by the cube referred to as Hybrid cube.

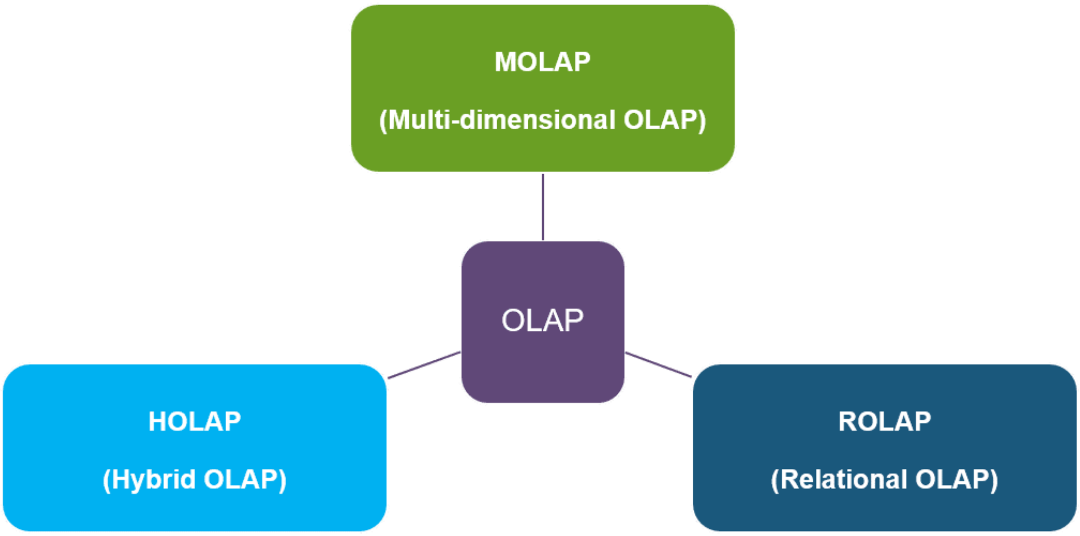
As per the formal definition, “Each cell within a multidimensional structure contains aggregated data related to elements along each of the dimensions.” The multidimensional analytical databases are helpful in providing data-related answers to complex business queries quickly and accurately. Further, unlike other data models, OLAP in data warehousing enables users to view data from different angles and dimensions, thereby presenting a broader analysis for business purposes.

It has been observed that the OLAP cubes answers a query in 0.1% of the time consumed for the similar query by an OLTP (Online Transaction Processing) relational database.

OLAP systems are mainly classified into three :

- **MOLAP (Multi-dimensional OLAP)**
- **ROLAP (Relational OLAP) :** works with relational databases

- **HOLAP (Hybrid OLAP):** database divides data between relational and specialized storage



Basic operations of OLAP

BASIC CONCEPT OF ASSOCIATION ANALYSIS AND CLUSTER ANALYSIS

Association Analysis

Association mining aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items or objects in transaction databases, relational database or other data repositories. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control, cross-marketing, catalog design, loss-leader analysis, clustering, classification, etc.

EXAMPLES:

Rule Form: Body->Head [Support, confidence]

Buys (X, “Computer”) -> Buys (X, “Software”) [40%, 50%]

ASSOCIATION RULE:

BASIC CONCEPTS:

Given: (1) database of transaction, (2) each transaction is a list of items (purchased by a customer in visit)

Find: all rules that correlate the presence of one set of items with that of another set of items.
• E.g., 98% of people who purchase tires and auto accessories also get automotive— services done.

• E.g., Market Basket Analysis — This process analyzes customer buying habits by finding associations between the different items that customers place in their “Shopping Baskets”.

The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customer.

APPLICATIONS:

Maintenance agreement (what the store should do to boost maintenance agreement sales)
Home Electronics (what other products should the store stocks up?)
Attached mailing in direct marketing

ASSOCIATION RULE:

An association rule is an implication expression of the form X◊Y, where X and Y are disjoint itemsets, i.e., $X \cap Y = \emptyset$. The strength of an association rule can be measured in terms of its support and confidence. Support determines how often a rule is applicable to a given data set, while confidence determines how frequently items in Y appear in transactions that contain X. The formal definition of these metrics are

Support, s(X->Y) = $\sigma(X \cup Y)$ N

Confidence, c(X->Y) =$\sigma(X \cup Y)$ $\sigma(X)$

Cluster Analysis

- The process of **partitioning** a set of data objects (or observations) into subsets (clusterS).
- Similar objects in a same cluster,
- Objects in different clusters are supposed to be different.

Clustering is known as **unsupervised learning** because the class label information is not present. For this reason, clustering is a form of **learning by observation**, rather than learning by examples.

Requirements for Cluster Analysis

1. **Scalability** (high)
   Clustering on only a sample of a given large data set may lead to biased results
2. **Ability to deal with different types of attributes**
   e.g. graphs, sequences, images, and documents.
3. **Discovery of clusters with arbitrary shape** :
   a cluster could be of any shape.
4. **Requirements for domain knowledge to determine input parameters**
   It’s hard to determine the parameter
5. **Deal with noisy data** (**Outliers**)
   Need clustering methods that are robust to noise.
6. **Incremental clustering**: incremental update, avoid recomputing a new clustering from scratch
   insensitive to input order: the change of input order doesn’t change output

7. **Capability of clustering high-dimensionality data**
   Finding clusters of data objects in a high- dimensional space is challenging, especially considering that such data can be very sparse and highly skewed.
8. **Constraint-based**
9. **Interpretability and usability**
   It is important to study how an application goal may influence the selection of clustering features and clustering methods.

Clustering Methods

1. **The partitioning criteria**
   e.g. hierarchical or not
2. **Separation of clusters**
   e.g. clusters are mutually exclusive or not
3. **Similarity measure**
   e.g.
   - distance
     - often take advantage of optimization techniques
     - e.g. Euclidean space, road network, vector space,
   - connectivity based on density or continuity
     - can often find clusters of arbitrary shape
4. **Clustering space**
   search for clusters within the entire given data space or subspace

   Subspace clustering discovers clusters and subspaces (often of low dimensionality) that manifest object similarity.

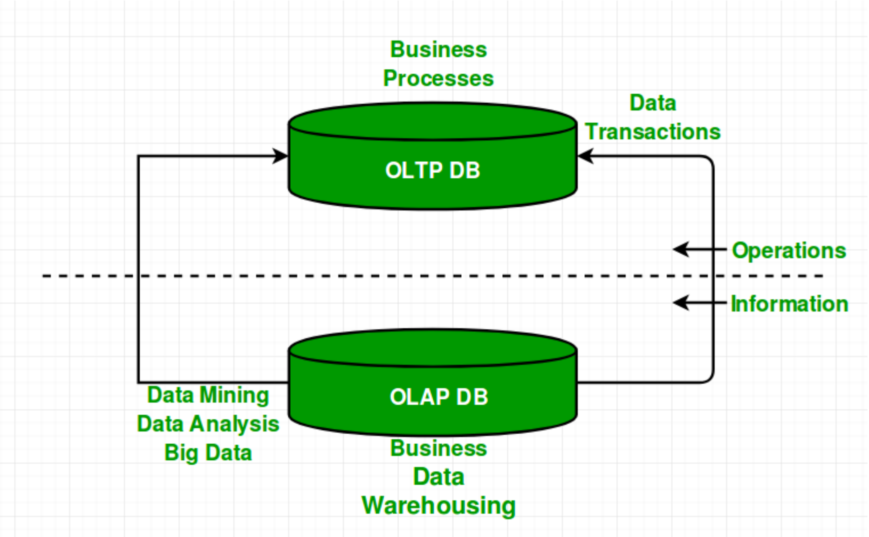
TRANSACTION PROCESSING V/S ANALYTIC PROCESSING

Transaction processing:

- Each transaction involves a relatively small amount of data
- There are inserts and updates to one or more tables
- The database should be normalized, ie, any piece of information should be in one place only, with a very few exceptions
- There are often requirements for audit data: who created the transaction when
- The data typically requires validation checks before processing (valid customer, product, account #, etc)

Analytical Processing

- Read-only, unless you need to build a temporary table, or populate a results table for multiple reports
- Often large volumes of data
- Database may be denormalized for faster performance
- No validations required unless the source transaction system has been sloppy

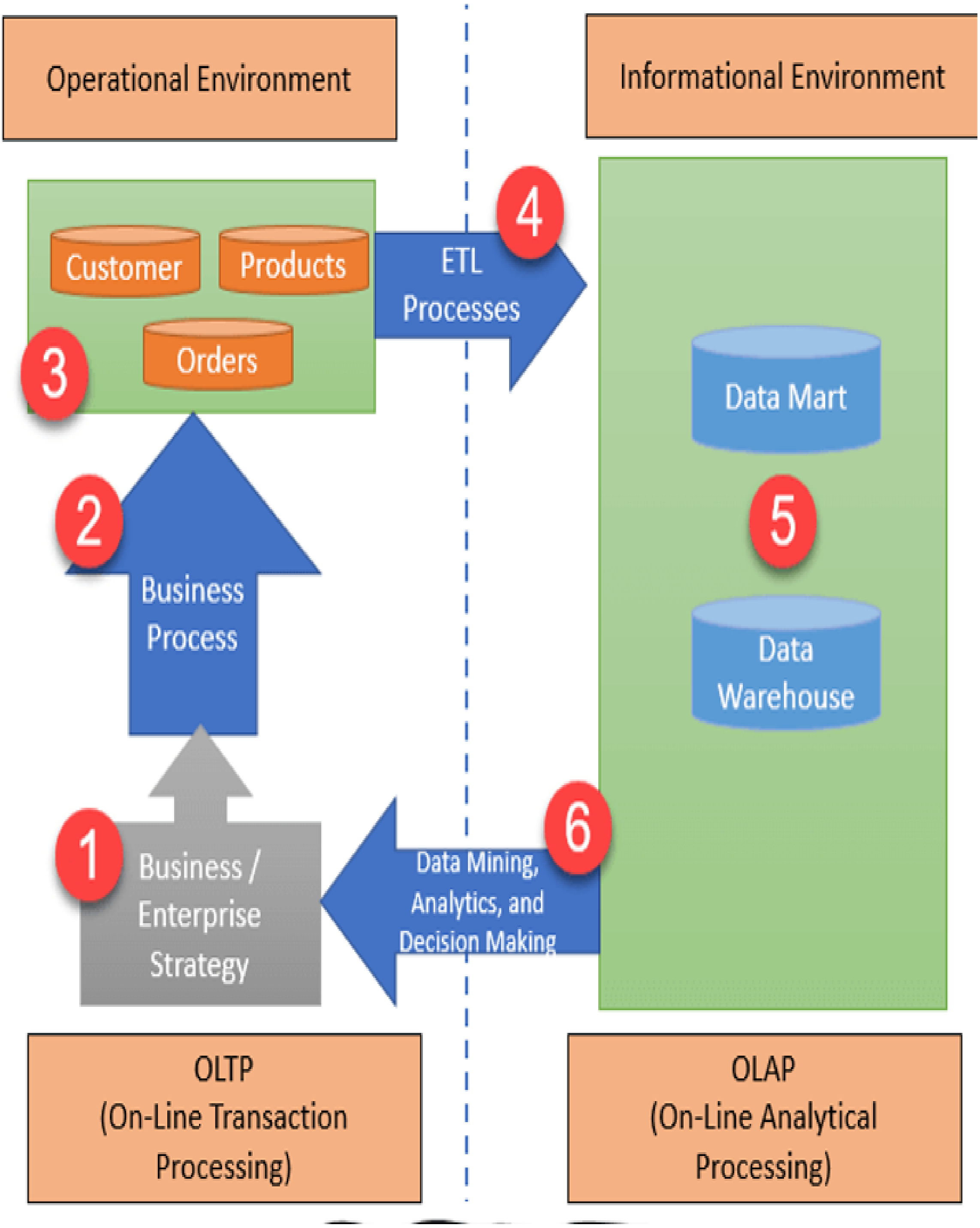


Transactional processing and Analytical Processing

OLTP V/S OLAP

OLAP stands for On-Line Analytical Processing. It is used for analysis of database information from multiple database systems at one time such as sales analysis and forecasting, market research, budgeting and etc. Data Warehouse is the example of OLAP system.

OLTP stands for On-Line Transactional processing. It is used for maintaining the online transaction and record integrity in multiple access environments. OLTP is a system that manages very large number of short online transactions for example, ATM.

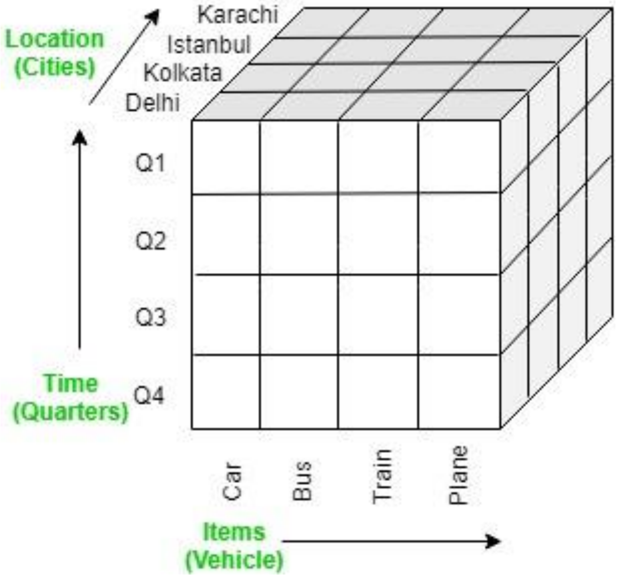

Operational Environment
Informational Environment
Customer
Products
Orders
3
4
ETL Processes
Data Mart
5
Data Warehouse
2
Business Process
1
Business / Enterprise Strategy
6
Data Mining, Analytics, and Decision Making
OLTP (On-Line Transaction Processing)
OLAP (On-Line Analytical Processing)

Architecture of OLTP and OLAP

Sr. No.	Key	OLAP	OLTP
1	Basic	It is used for data analysis	It is used to manage very large number of online short transactions
2	Database Type	It uses data warehouse	It uses traditional DBMS
3	Data Modification	It manages all insert, update and delete transaction	It is mainly used for data reading
4	Response time	Processing is little slow	In Milliseconds
5	Normalization	Tables in OLAP database are not normalized.	Tables in OLTP database are normalized.

OLAP Operations

OLAP stands for ***Online Analytical Processing*** Server. It is a software technology that allows users to analyze information from multiple database systems at the same time. It is based on multidimensional data model and allows the user to query on multi-dimensional data (eg. Delhi -> 2018 -> Sales data). OLAP databases are divided into one or more cubes and these cubes are known as *Hyper-cubes*.

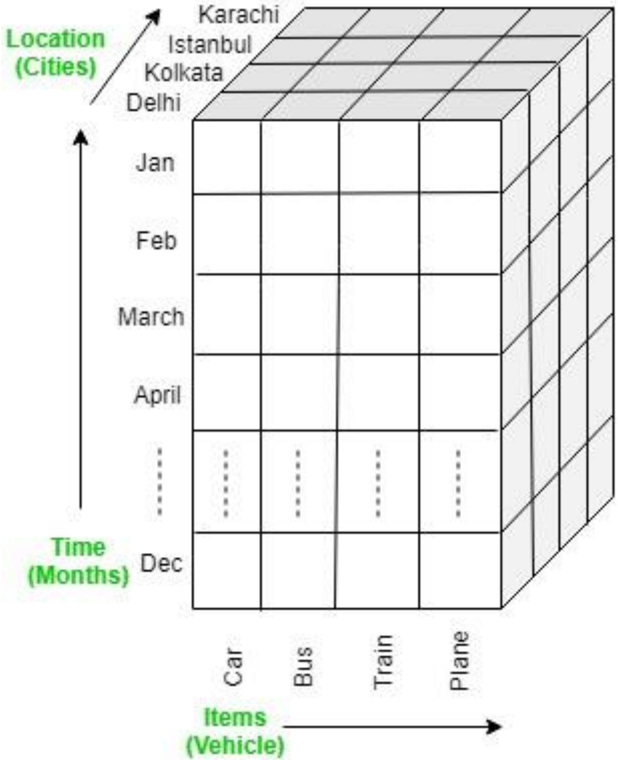


OLAP operations:

There are five basic analytical operations that can be performed on an OLAP cube:

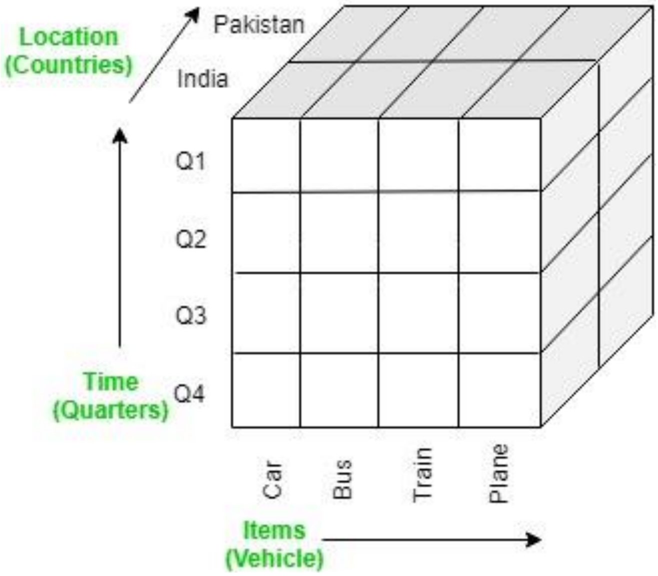
1. **Drill down:** In drill-down operation, the less detailed data is converted into highly detailed data. It can be done by:
   - Moving down in the concept hierarchy
   - Adding a new dimension

   In the cube given in overview section, the drill down operation is performed by moving down in the concept hierarchy of *Time* dimension (Quarter -> Month).

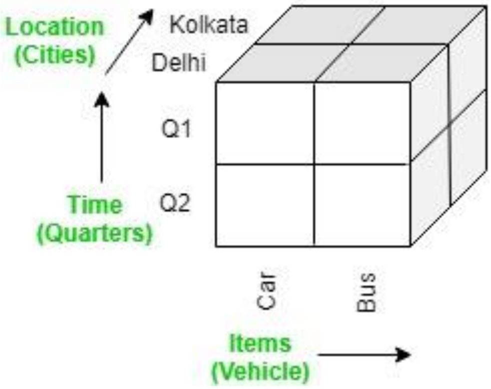


2. **Roll up:** It is just opposite of the drill-down operation. It performs aggregation on the OLAP cube. It can be done by:
   - Climbing up in the concept hierarchy
   - Reducing the dimensions

In the cube given in the overview section, the roll-up operation is performed by climbing up in the concept hierarchy of *Location* dimension (City -> Country).

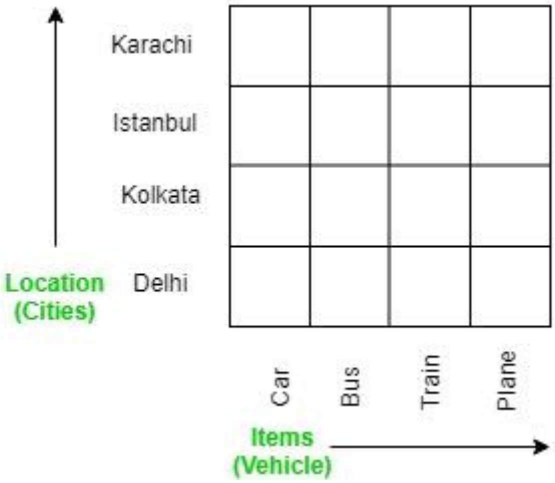


3. **Dice:** It selects a sub-cube from the OLAP cube by selecting two or more dimensions. In the cube given in the overview section, a sub-cube is selected by selecting following dimensions with criteria:
   - Location = “Delhi” or “Kolkata”
   - Time = “Q1” or “Q2”
   - Item = “Car” or “Bus”

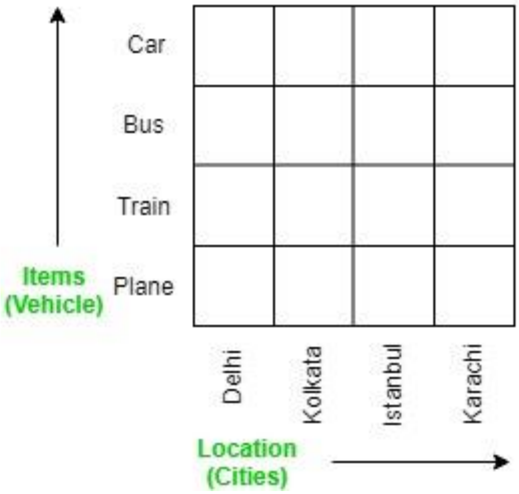


4. **Slice:** It selects a single dimension from the OLAP cube which results in a new sub-cube creation. In the cube given in the overview section, Slice is performed on the

dimension Time = “Q1”.

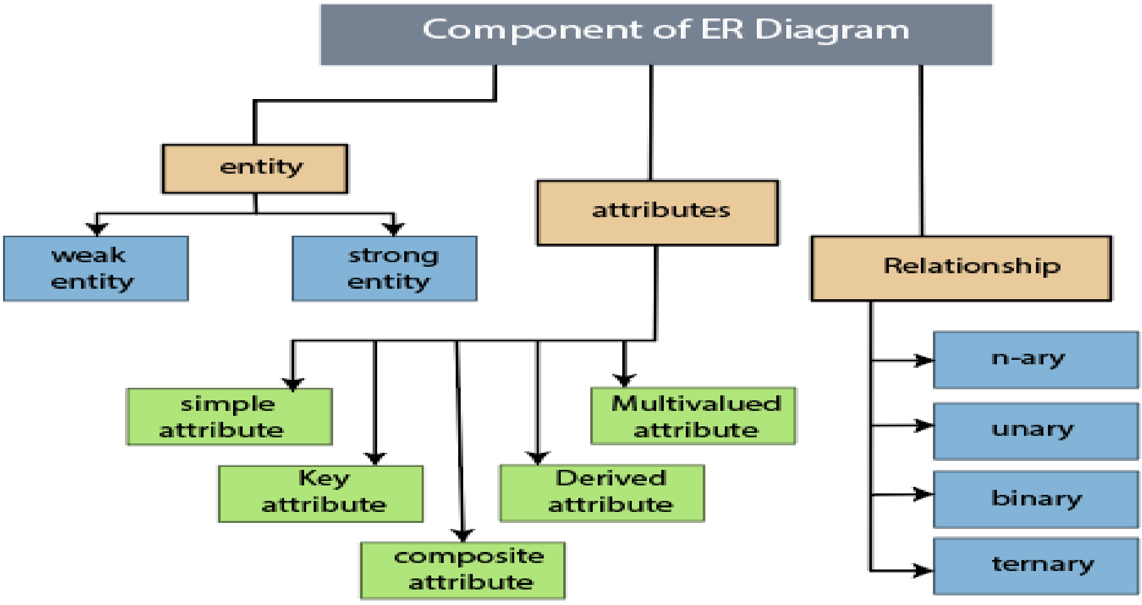


5. **Pivot:** It is also known as *rotation* operation as it rotates the current view to get a new view of the representation. In the sub-cube obtained after the slice operation, performing pivot operation gives a new view of it.



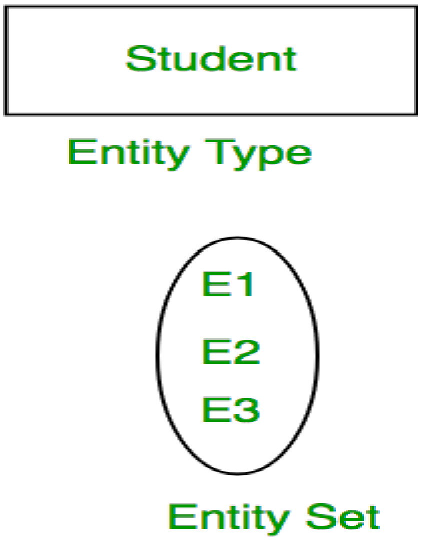
DATA MODELS FOR OLTP (ER MODEL)

ER Model is used to model the logical view of the system from data perspective which consists of these components:

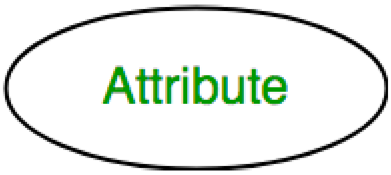


Entity, Entity Type, Entity Set –
An Entity may be an object with a physical existence – a particular person, car, house, or employee – or it may be an object with a conceptual existence – a company, a job, or a university course.

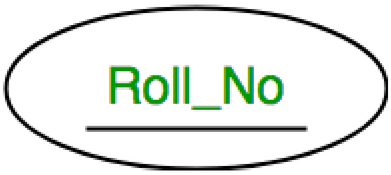
An Entity is an object of Entity Type and set of all entities is called as entity set. e.g.; E1 is an entity having Entity Type Student and set of all students is called Entity Set. In ER diagram, Entity Type is represented as:



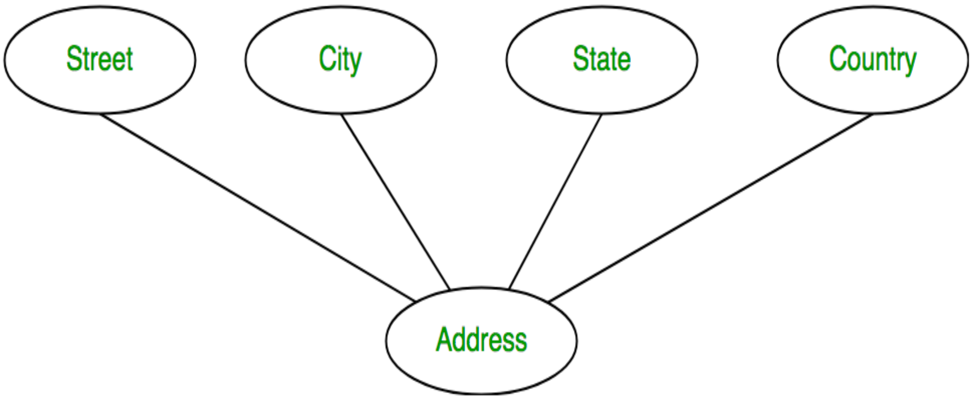
Attribute(s):
Attributes are the **properties which define the entity type**. For example, Roll_No, Name, DOB, Age, Address, Mobile_No are the attributes which defines entity type Student. In ER diagram, attribute is represented by an oval.



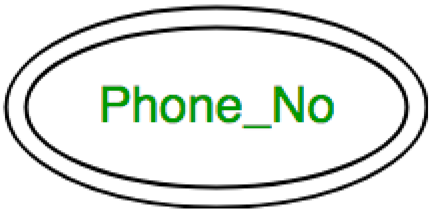
1. Key Attribute –
The attribute which **uniquely identifies each entity** in the entity set is called key attribute.For example, Roll_No will be unique for each student. In ER diagram, key attribute is represented by an oval with underlying lines.



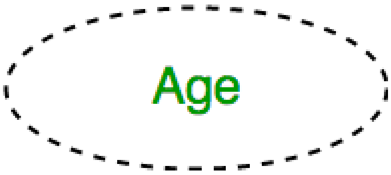
2. Composite Attribute –
An attribute **composed of many other attribute** is called as composite attribute. For example, Address attribute of student Entity type consists of Street, City, State, and Country. In ER diagram, composite attribute is represented by an oval comprising of ovals.



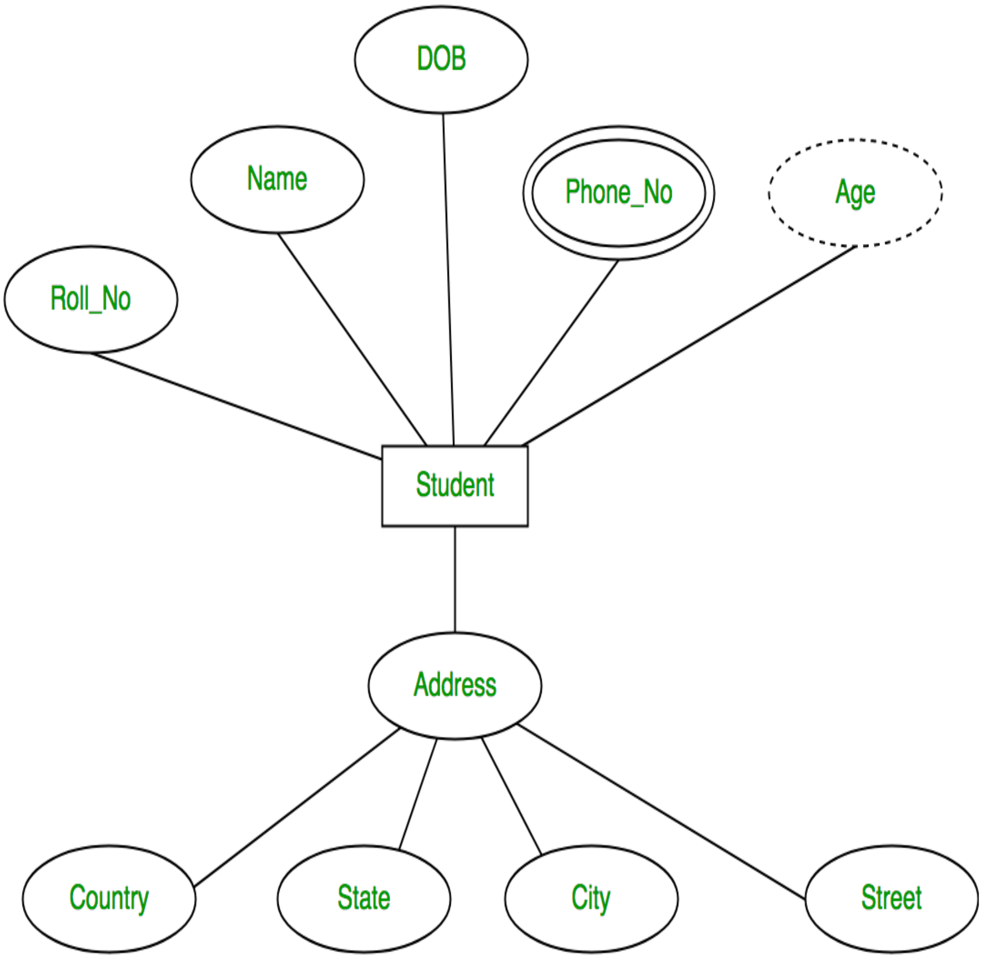
3. Multivalued Attribute –
An attribute consisting **more than one value** for a given entity. For example, Phone_No (can be more than one for a given student). In ER diagram, multivalued attribute is represented by double oval.



4. Derived Attribute –
An attribute which can be **derived from other attributes** of the entity type is known as derived attribute. e.g.; Age (can be derived from DOB). In ER diagram, derived attribute is represented by dashed oval.

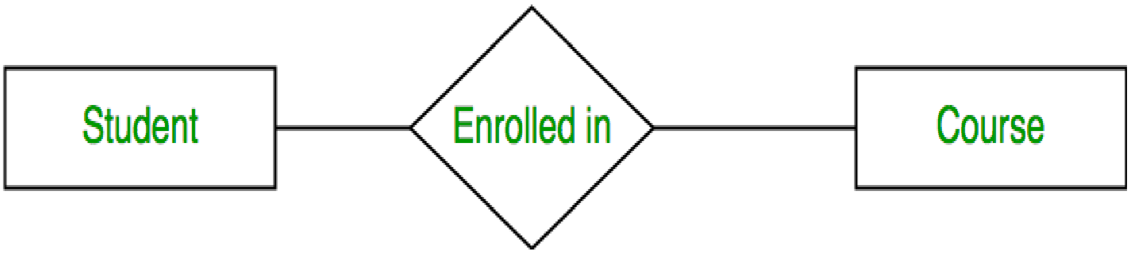


The complete entity type **Student** with its attributes can be represented as:

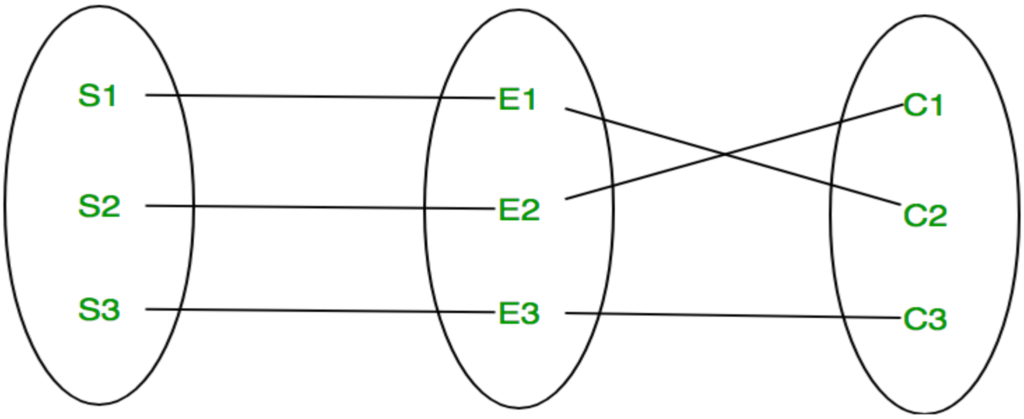


Relationship Type and Relationship Set:
A relationship type represents the **association between entity types**. For example,‘Enrolled

in’ is a relationship type that exists between entity type Student and Course. In ER diagram, relationship type is represented by a diamond and connecting the entities with lines.

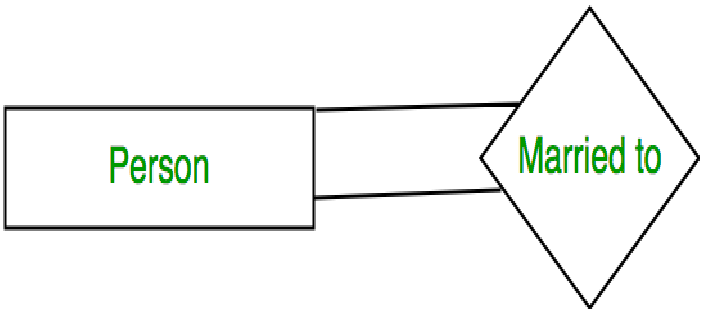


A set of relationships of same type is known as relationship set. The following relationship set depicts S1 is enrolled in C2, S2 is enrolled in C1 and S3 is enrolled in C3.

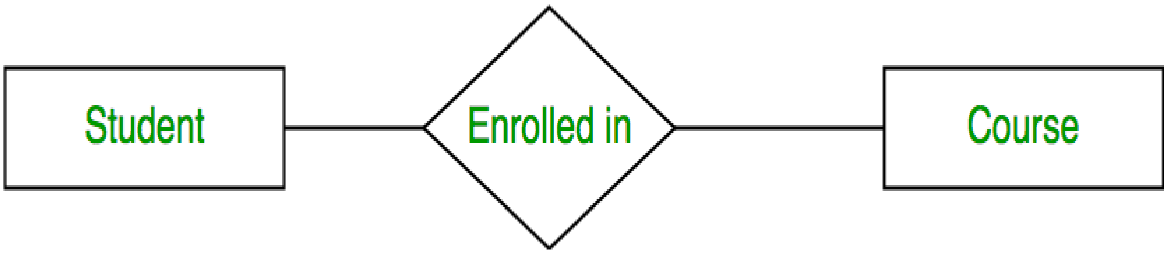


Degree of a relationship set:
The number of different entity sets **participating in a relationship** set is called as degree of a relationship set.

1. Unary Relationship –
When there is **only ONE entity set participating in a relation**, the relationship is called as unary relationship. For example, one person is married to only one person.

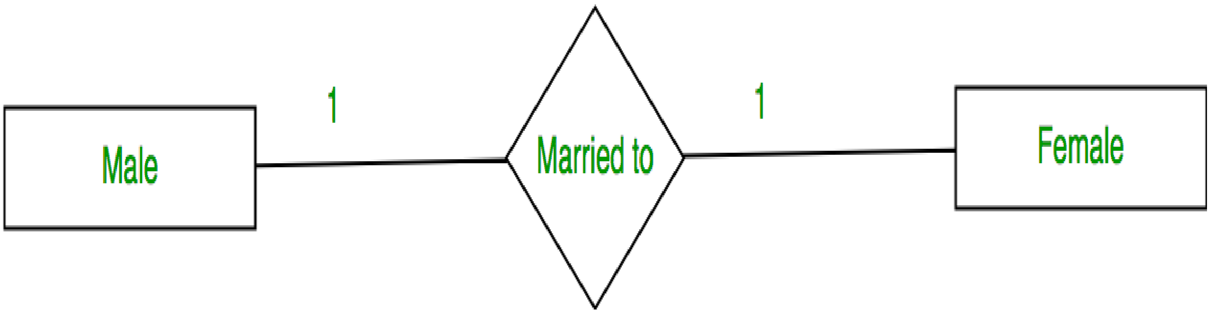


2. Binary Relationship –
When there are **TWO entities set participating in a relation**, the relationship is called as binary relationship.For example, Student is enrolled in Course.

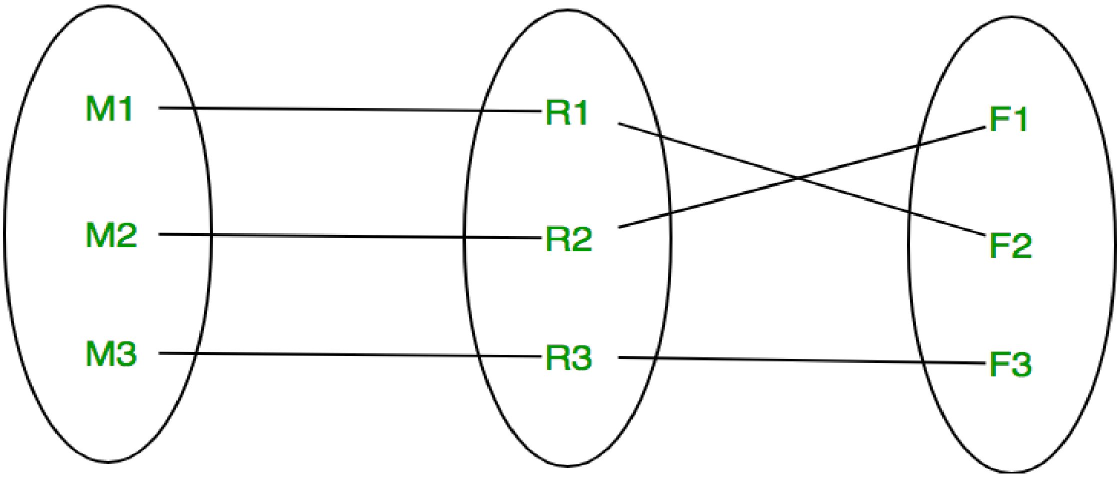


3. n-ary Relationship –
When there are n entities set participating in a relation, the relationship is called as n-ary relationship.

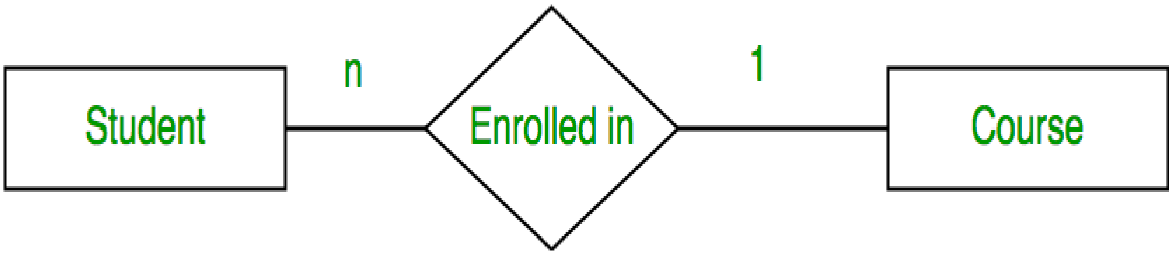
Cardinality:
The **number of times an entity of an entity set participates in a relationship** set is known as cardinality. Cardinality can be of different types:
1. One to one – When each entity in each entity set can take part **only once in the relationship**, the cardinality is one to one. Let us assume that a male can marry to one female and a female can marry to one male. So the relationship will be one to one.



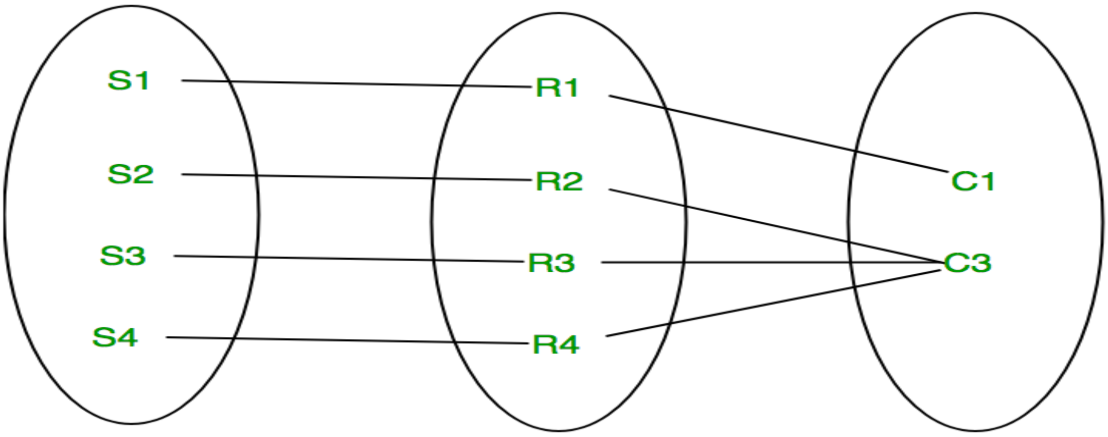
Using Sets, it can be represented as:



2. Many to one – When entities in one entity set **can take part only once in the relationship set and entities in other entity set can take part more than once in the relationship set,** cardinality is many to one. Let us assume that a student can take only one course but one course can be taken by many students. So the cardinality will be n to 1. It means that for one course there can be n students but for one student, there will be only one course.

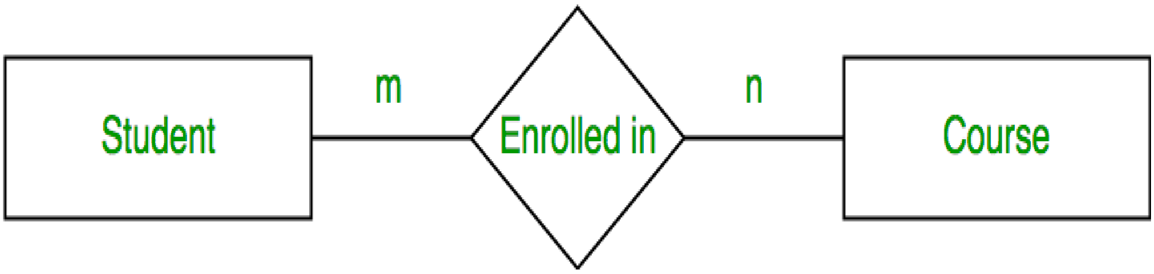


Using Sets, it can be represented as:

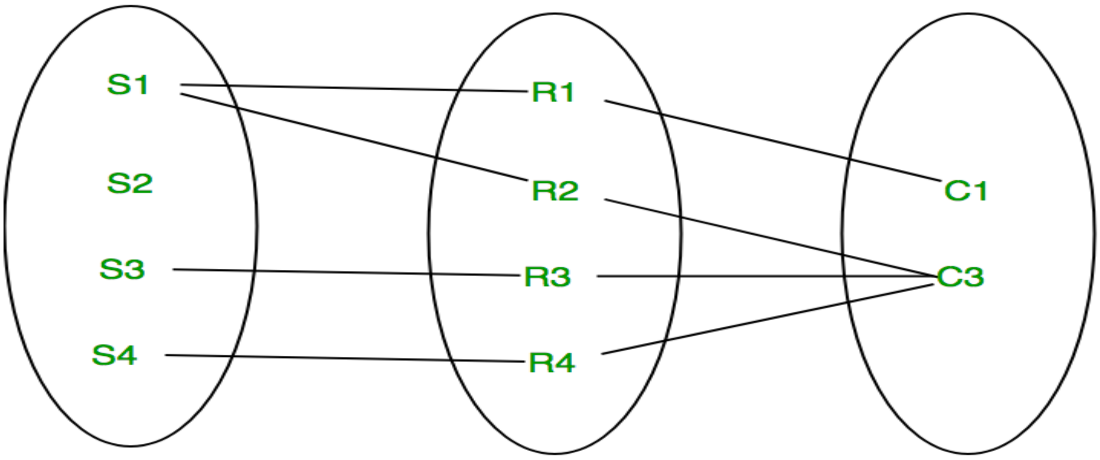


In this case, each student is taking only 1 course but 1 course has been taken by many students.

3. Many to many – When entities in all entity sets can **take part more than once in the relationship** cardinality is many to many. Let us assume that a student can take more than one course and one course can be taken by many students. So the relationship will be many to many.



Using sets, it can be represented as:



In this example, student S1 is enrolled in C1 and C3 and Course C3 is enrolled by S1, S3 and S4. So it is many to many relationships.

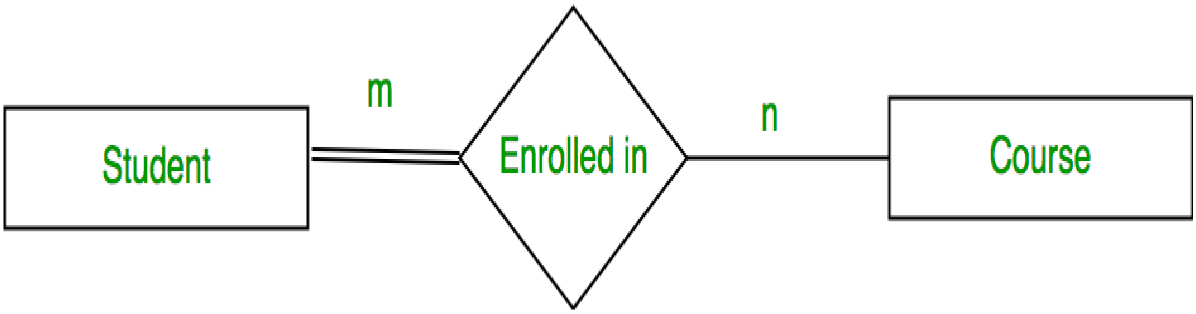
Participation Constraint:

Participation Constraint is applied on the entity participating in the relationship set.

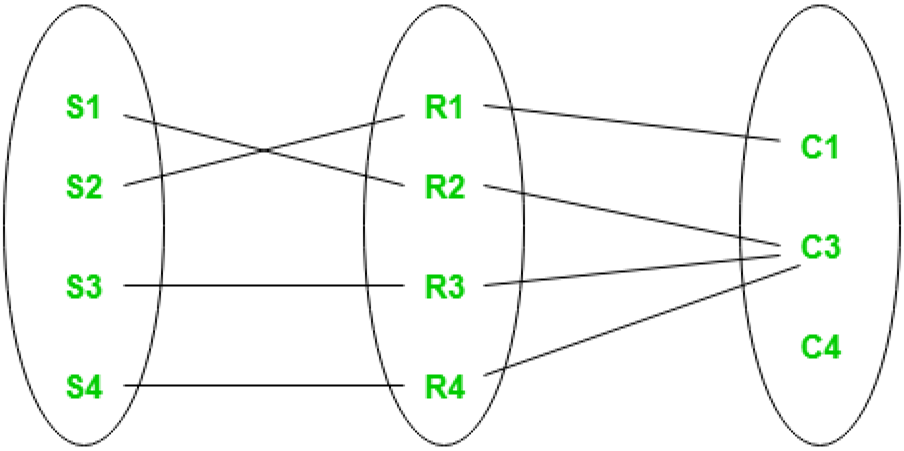
1. Total Participation – Each entity in the entity set **must participate** in the relationship. If each student must enroll in a course, the participation of student will be total. Total participation is shown by double line in ER diagram.

2. Partial Participation – The entity in the entity set **may or may NOT participat**e in the relationship. If some courses are not enrolled by any of the student, the participation of course will be partial.

The diagram depicts the ‘Enrolled in’ relationship set with Student Entity set having total participation and Course Entity set having partial participation.



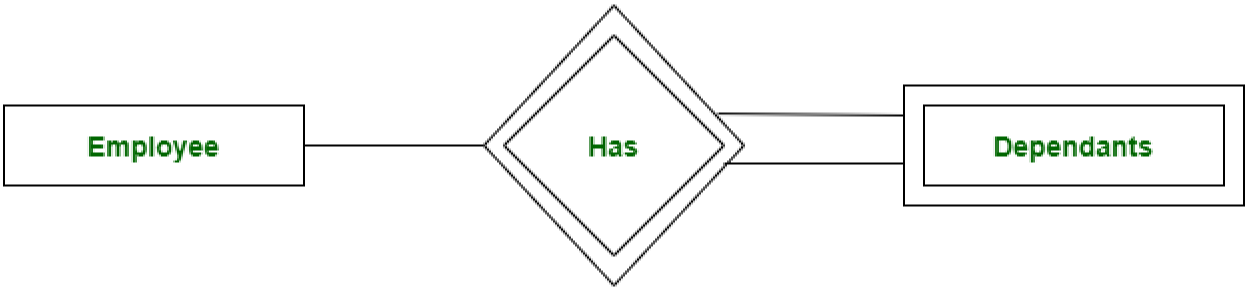
Using set, it can be represented as,

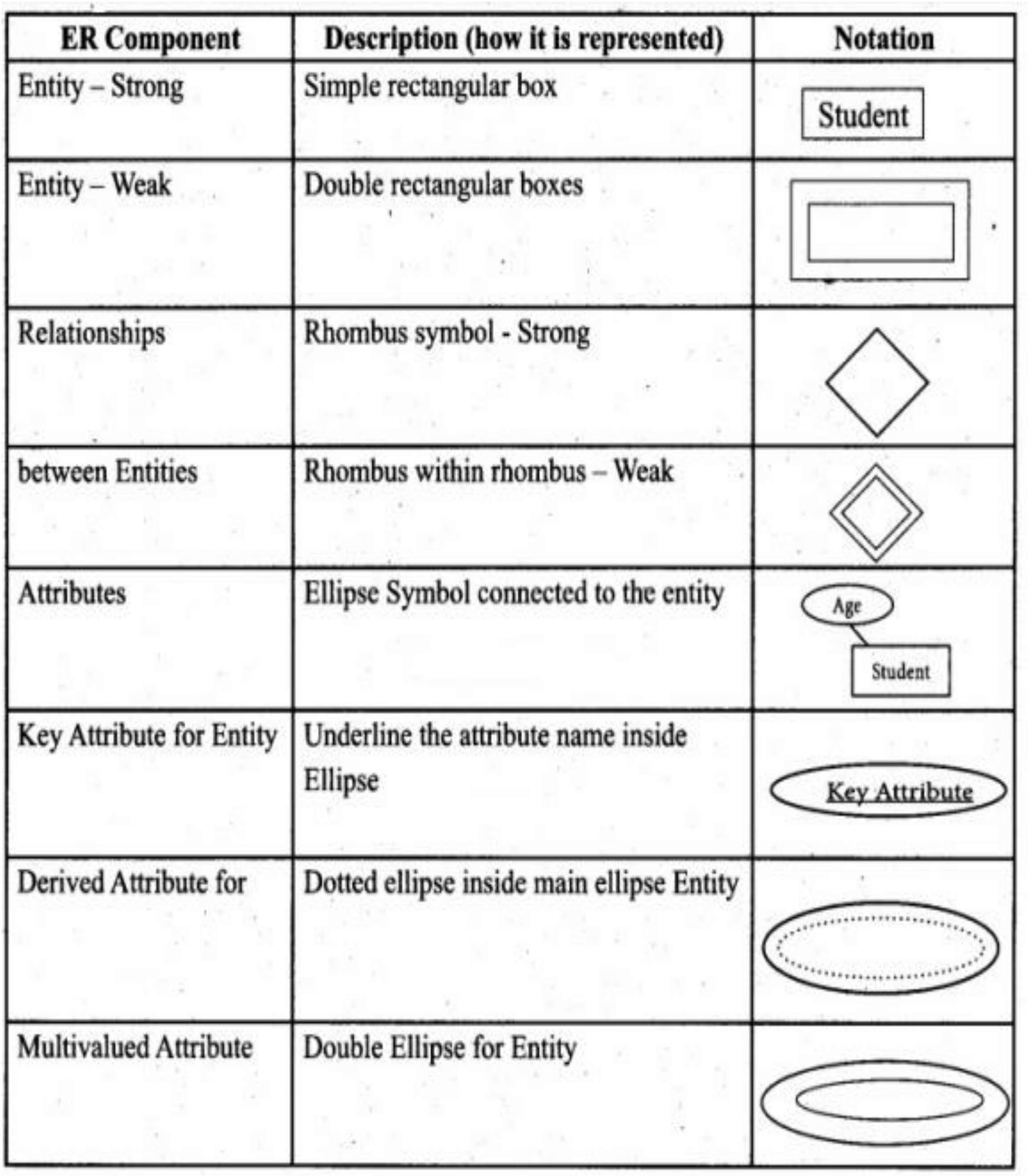


Every student in Student Entity set is participating in relationship but there exists a course C4 which is not taking part in the relationship.

Weak Entity Type and Identifying Relationship:
As discussed before, an entity type has a key attribute which uniquely identifies each entity in the entity set. But there exists **some entity type for which key attribute can't be defined**. These are called Weak Entity type.
For example, A company may store the information of dependents (Parents, Children, Spouse) of an Employee. But the dependents don't have existence without the employee. So Dependent will be weak entity type and Employee will be Identifying Entity type for Dependent.

A weak entity type is represented by a double rectangle. The participation of weak entity type is always total. The relationship between weak entity type and its identifying strong entity type is called identifying relationship and it is represented by double diamond.





ER Component	Description (how it is represented)	Notation
Entity – Strong	Simple rectangular box	Student
Entity – Weak	Double rectangular boxes	
Relationships	Rhombus symbol - Strong	
between Entities	Rhombus within rhombus – Weak	
Attributes	Ellipse Symbol connected to the entity	Age Student
Key Attribute for Entity	Underline the attribute name inside Ellipse	Key Attribute
Derived Attribute for	Dotted ellipse inside main ellipse Entity	
Multivalued Attribute	Double Ellipse for Entity	

Component description of ER Diagram

OLAP (STAR & SNOWFLAKE SCHEMA)

Star Schema vs. Snowflake Schema: The Main Difference

The two main elements of the dimensional model of the star and snowflake schema are:

1. **Facts table**. A table with the most considerable amount of data, also known as a **cube**.

2. **Dimension tables**. The derived data structure provides answers to ad hoc queries or dimensions, often called **lookup tables**.

Connecting chosen **dimensions** on a **facts table** forms the schema. Both the star and snowflake schemas make use of the dimensionality of data to model the storage system.

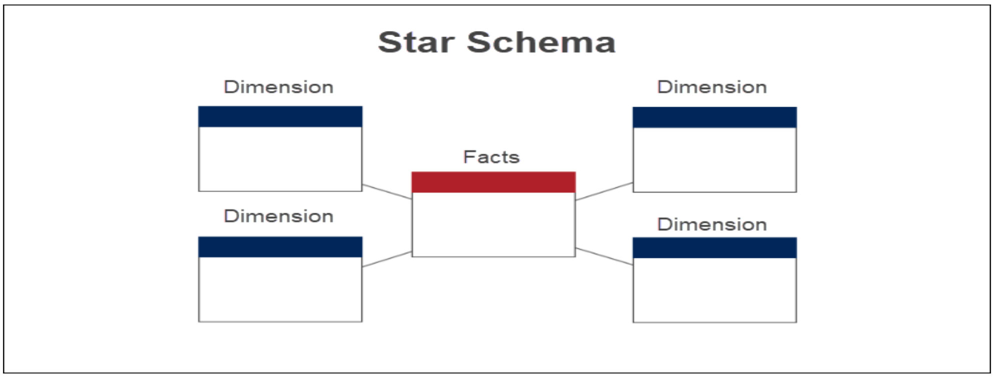
The main differences between the two schemas are:

	Star Schema	**Snowflake Schema**
Elements	Fact table Dimension tables	Fact table Dimension tables Subdimension tables
Structure	Star-shaped	Snowflake shaped
Dimensions	One table per dimension	Multiple tables for each dimension
Model Direction	Top-down	Bottom-up
Storage space	Uses more storage	Uses less space
Normalization	Denormalized dimension tables	Normalized dimension tables
Query Performance	Fast, fewer JOINs needed because of fewer foreign keys	Slow, more JOINs required because of more foreign keys
Query Complexity	Simple and easier to understand	Complicated and more challenging to understand
Data Redundancy	High	Low
Use case	Dimension tables with several rows, typical with data marts	Dimension tables with multiple rows found with data warehouses

Due to the complexity of the snowflake schema and the lower performances, the star schema is the preferred option whenever possible. One typical way to get around the problems in the snowflake schema is to decompose the dedicated storage into multiple smaller entities with a star schema.

What Is a Star Schema?

A star schema is a logical structure for the development of data marts and simpler data warehouses. The simple model consists of dimension tables connected to a facts table in the center.

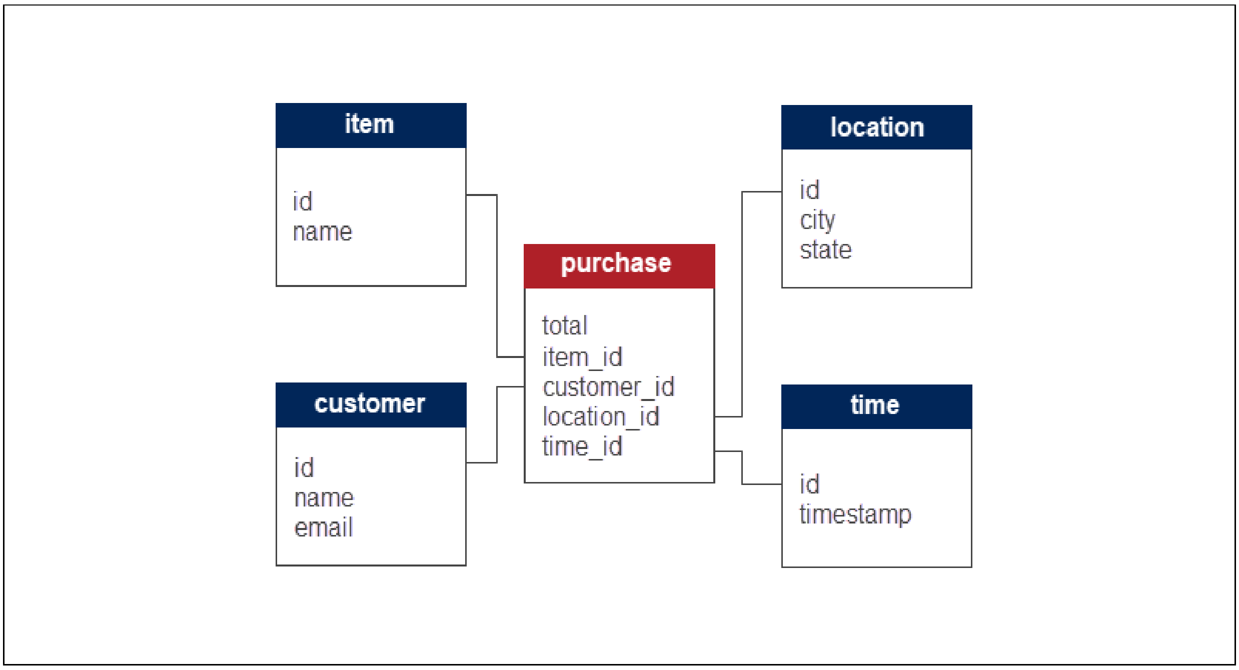


The facts table typically consists of:

- Quantifiable numerical data, such as values or counts.
- References to the dimensions through foreign keys.

The lookup tables represent descriptive information directly connected to the facts table.

For example, to model the sales of an ecommerce business, the facts table for purchases might contain the total price of the purchase. On the other hand, dimensional tables have descriptive information about the items, customer data, the time or location of purchase.



The star schema for the analysis of purchases in the example has four dimensions. The facts table connects to the dimensional tables through the concept of foreign and primary keys. Apart from the numerical data, the facts table therefore also consists of foreign keys to define relations between tables.

Characteristics of a Star Schema

The main characteristics of the star schema are:

- **Simplified and fast queries**. Fewer JOIN operations due to denormalization make information more readily available.
- **Simple relationships.** The schema works great with one-to-one or one-to-many relationships.
- **Singular dimensionality**. One table describes each dimension.
- **OLAP friendly**. OLAP systems widely use star schema to design data cubes.

Drawbacks of a Star Schema

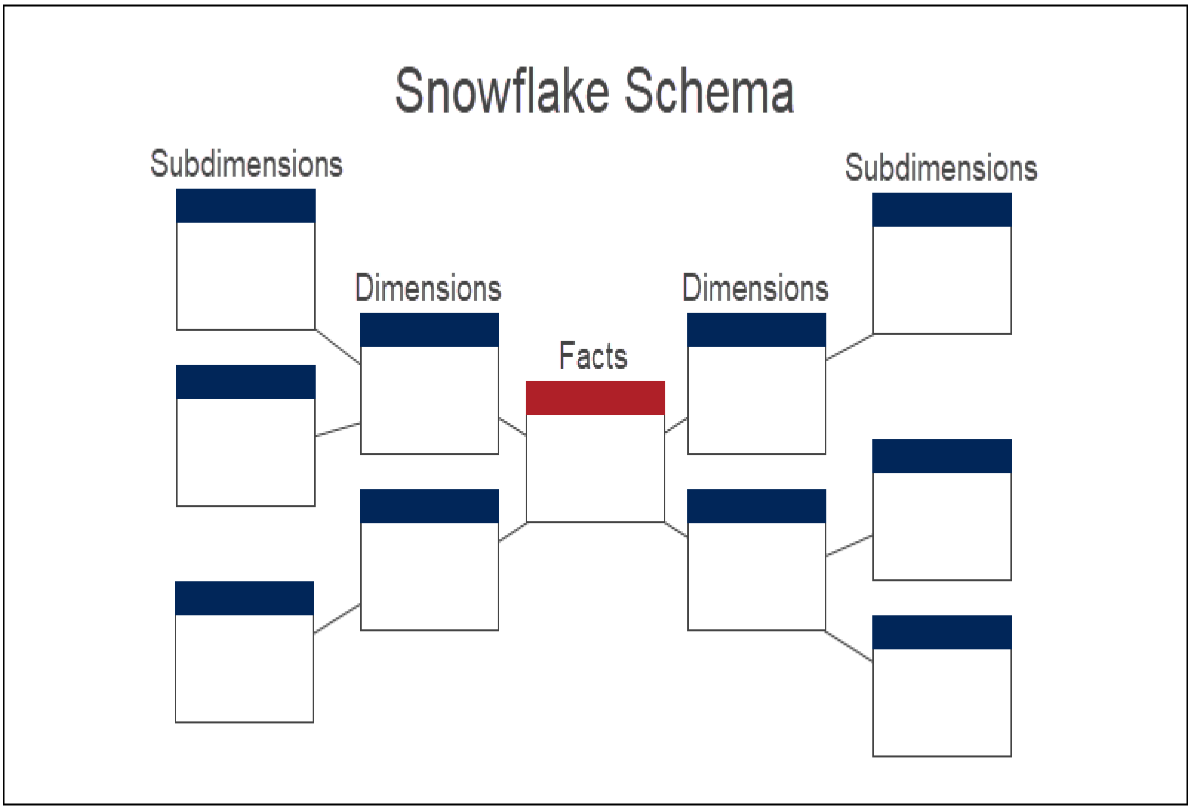
The disadvantages of using the star schema are:

- **Redundancy**. The dimensional tables are one-dimensional, and data redundancy is present.
- **Low integrity**. Due to denormalization, updating information is a complex task.
- **Limited queries**. The set of questions is limited, which also narrows down the analytical power.

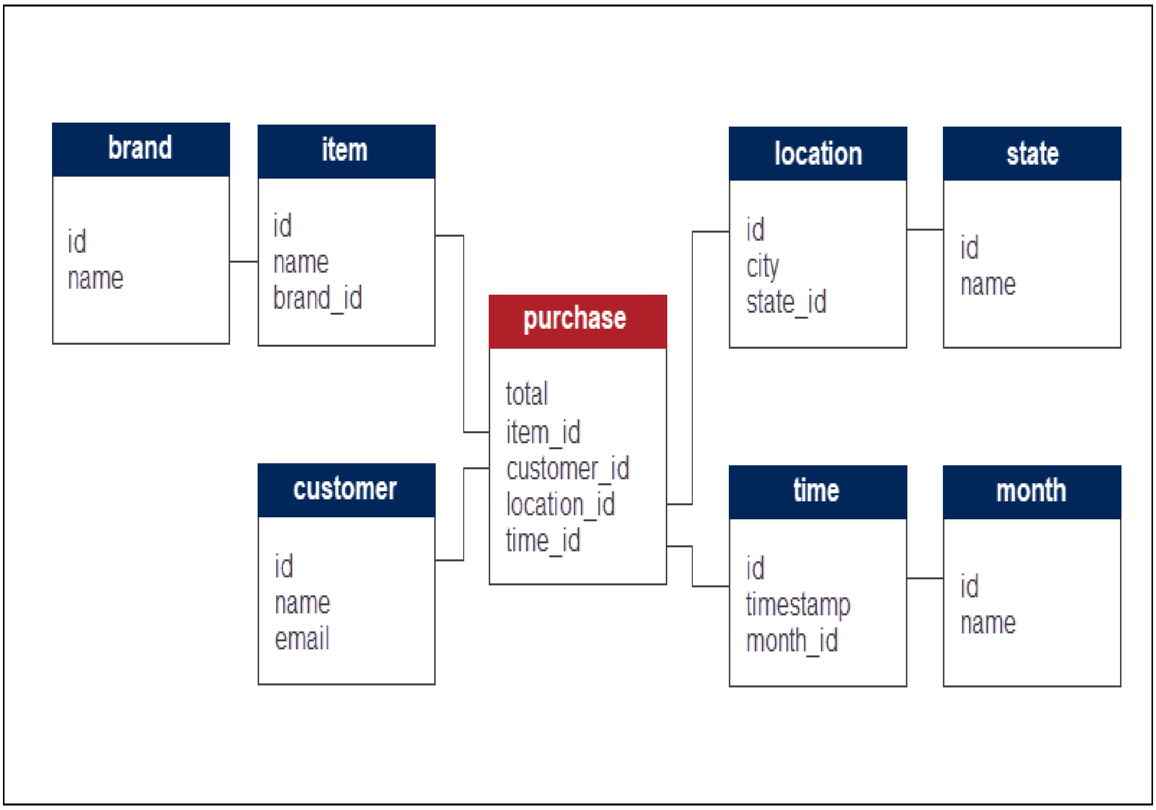
What Is a Snowflake Schema?

The snowflake schema has a branched-out logical structure used in large data warehouses. From the center to the edges, entity information goes from general to more specific.

Apart from the dimensional model's common elements, the snowflake schema further decomposes dimensional tables into subdimensions.



The ecommerce sales analysis model from the previous example further branches ("snowflakes") into smaller categories and subcategories of interest.



The four dimensions decompose into subdimensions. The lookup tables further normalize through a series of connected objects.

Characteristics of a Snowflake Schema

The main features of the snowflake schema include:

- **Small storage**. The snowflake schema does not require as much storage space.
- **High granularity**. Dividing tables into subdimensions allows analysis at various depths of interest. Adding new subdimensions is a simple process as well.
- **Integrity**. Due to normalization, the schema has a higher level of data integrity and low redundancies.

Drawbacks of a Snowflake Schema

The weaknesses of the snowflake schema are:

- **Complexity**. The database model is complex, and so are the executed queries. Multiple multidimensional tables make the design complicated to work with overall.
- **Slow processing**. Many lookup tables require multiple JOIN operations, which slows down information retrieval.
- **Hard to maintain**. A high level of granularity makes the schema hard to manage and maintain.

PART-A

1.List out the two categories of data mining tasks.

2. Discuss on the clustering methods.

3. Define Entity.

4.List out the characteristics of a star schema.

5.What are the two main elements of the dimensional model in star and snowflake schema?

6. State the different types of cardinality.

7. List out the drawbacks of snowflake schema.

8. Discuss on Analytical processing.

9. Discuss on transactional processing.

10.State the requirements for cluster analysis

PART-B

1.Discuss in detail on the various Data Mining Techniques.

2.Explain in detail about the various OLTP operations.

3. Explain the concept of Association Analysis and clustering Analysis.

4. Differentiate between OLTP and OLAP.

5.Describe in detail about the ER model with suitable diagrams.

6.Discuss in detail about the star and snowflake schema with suitable diagrams.