UNIT: 9 Fundamentals of Database

Contents:

❑Introduction to DBMS,

❑DBMS Models, Database Design and Data Security,

❑ Data Warehouse and Data Mining, Big Data

❑Database Administrator

What is Database?

- Database is a collection of related data and data is a collection of facts and figures that can be processed to produce information.

- In computing, a database is an organized collection of data stored and accessed electronically from a computer system

- A **database** is an organized collection of structured information, or data, typically stored electronically in a computer system.

- A **database** is usually controlled by a **database** management system (DBMS).

- The data can then be easily accessed, managed, modified, updated, controlled, and organized.

Database Management System

- A database management system stores data in such a way that it becomes easier to retrieve, manipulate, and produce information

- A collection of programs that enables us to store, modify, and extract information from a database.

- There are many different types of DBMSs, ranging from small systems that run on personal computers to huge systems that run on mainframes.

- The following are examples of database applications:
    - computerized library systems
    - automated teller machines
    - flight reservation systems
    - computerized parts inventory systems

Database Processing

- Data processing is **a series of operations that use information to produce a result**.

- Common data processing operations include validation, sorting, classification, calculation, interpretation, organization and transformation of data.

Six stages of data processing
Data collection. Collecting **data** is the first **step in data processing**. ...
Data preparation. Once the **data** is collected, it then enters the **data** preparation stage.
...
Data input. ...
Processing. ...
Data output/interpretation. ...
Data storage.

File Processing

- **Sequential file processing:**
  - The Sequential file organization is **a popular file organization in** the database management system (DBMS).
  - It is a simple technique for file organization structure.
  - The File organization in DBMS supports various data operations such as insert, update, delete, and retrieve the data.
  - It contains a fast and efficient method for the huge amount of data.
  - In this method, **files** can be easily stored in cheaper storage mechanism like magnetic tapes.
  - It is simple in design.
  - It requires no much effort to store the data.
  - Sequential file processing is used in situations where data can be processed in batches.
  - Payroll processing is a classic example of sequential processing.

File Processing Cont..

- **Direct-access file processing:**
  - There are many ways of organizing a file for direct-access.
  - First, the file must be stored on a direct-device like a disk, so that the records need not be processed in sequence.
  - Second, some means must be developed for determining the location of a particular record. Indexes are on common means

Characteristics of DBMS

Real-world entity:

- A modern DBMS is more realistic and uses real-world entities to design its architecture. It uses the behavior and attributes too. For example, a school database may use students as an entity and their age as an attribute.

Relation-based tables:

- DBMS allows entities and relations among them to form tables. A user can understand the architecture of a database just by looking at the table names.

Isolation of data and application:

- A database system is entirely different than its data. A database is an active entity, whereas data is said to be passive, on which the database works and organizes.
- DBMS also stores metadata, which is data about data, to ease its own process.

Less redundancy:

- DBMS follows the rules of normalization, which splits a relation when any of its attributes is having redundancy in values.
- Normalization is a mathematically rich and scientific process that reduces data redundancy.

Characteristics of DBMS Cont…

Consistency:

- Consistency is a state where every relation in a database remains consistent. There exist methods and techniques, which can detect attempt of leaving database in inconsistent state.

- A DBMS can provide greater consistency as compared to earlier forms of data storing applications like file-processing systems.

Query Language:

- DBMS is equipped with query language, which makes it more efficient to retrieve and manipulate data. A user can apply as many and as different filtering options as required to retrieve a set of data. Traditionally it was not possible where file-processing system was used

ACID Properties:

- DBMS follows the concepts of Atomicity, Consistency, Isolation, and Durability (normally shortened as ACID). These concepts are applied on transactions, which manipulate data in a database.

- ACID properties help the database stay healthy in multi-transactional environments and in case of failure.

Characteristics of DBMS Cont…

Multiuser and Concurrent Access:

- DBMS supports multi-user environment and allows them to access and manipulate data in parallel. Though there are restrictions on transactions when users attempt to handle the same data item, but users are always unaware of them.

Multiple views:

- DBMS offers multiple views for different users. A user who is in the Sales department will have a different view of database than a person working in the Production department.
- This feature enables the users to have a concentrate view of the database according to their requirements.

Security:

- Features like multiple views offer security to some extent where users are unable to access data of other users and departments. DBMS offers methods to impose constraints while entering data into the database and retrieving the same at a later stage.
- DBMS offers many different levels of security features, which enables multiple users to have different views with different features
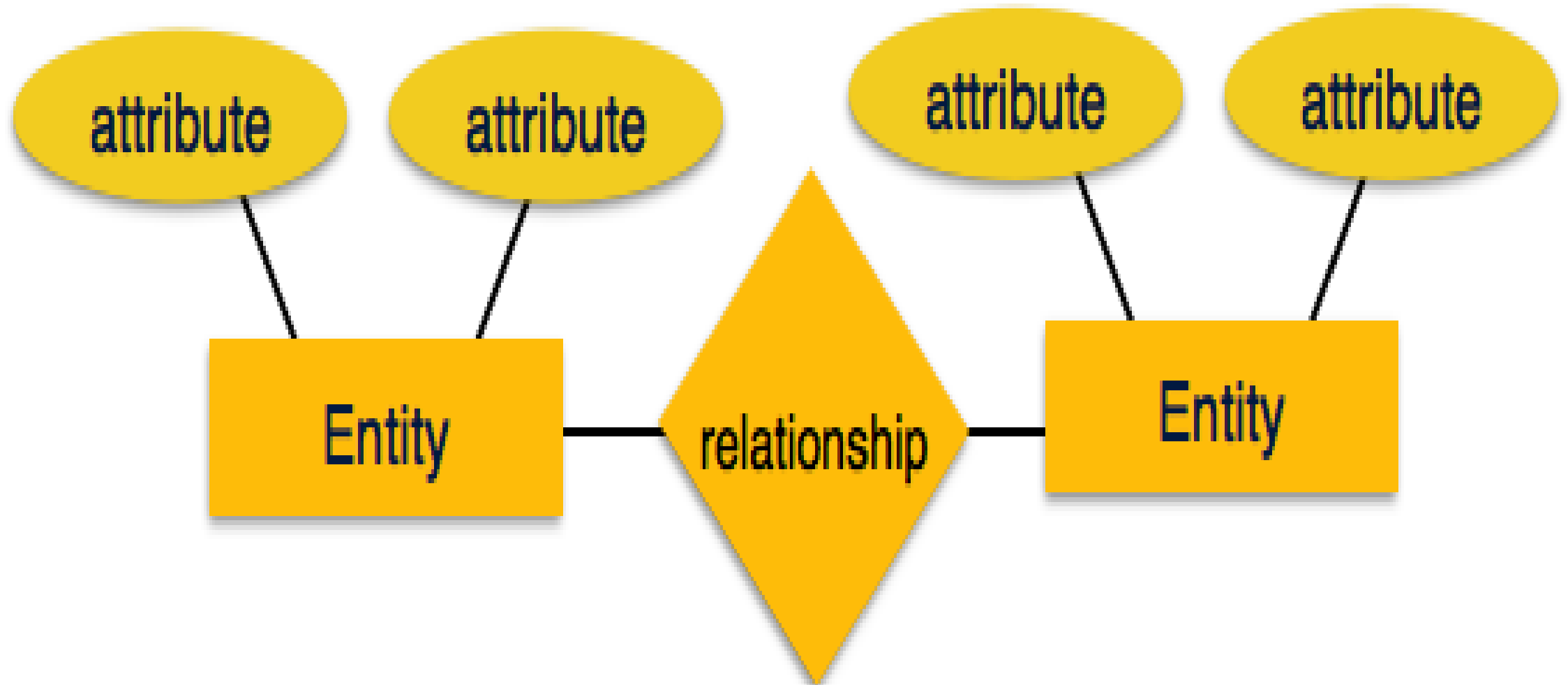
Database Models

- Data models define how the logical structure of a database is modeled.

- Data Models are fundamental entities to introduce abstraction in a DBMS.

- Data models define how data is connected to each other and how they are processed and stored inside the system.

- Two types:

1. Entity-Relationship Model

- Entity-Relationship (ER) Model is based on the notion of real-world entities and relationships among them.

- While formulating real-world scenario into the database model, the ER Model creates entity set, relationship set, general attributes, and constraints.

- ER model stands for an Entity-Relationship model. It is a high-level data model.

- This model is used to define the data elements and relationship for a specified system.

The E-R Model Cont...

CSIT IIT UNIT- 9 Fundamentals of Database   3/25/2023

The E-R Model Cont…

- ER Model is best used for the conceptual design of a database.
- ER Model is based on:

  • **Entities and their attributes**

- An entity in an ER Model is a real-world entity having properties called attributes.
- Every attribute is defined by its set of values called domain
- For example, in a school database, a student is considered as an entity. Student has various attributes like name, age, class, etc.
- Relationships among entities The logical association among entities is called relationship.

Entity

- Entities are represented by means of rectangles. Rectangles are named with the entity set they represent.
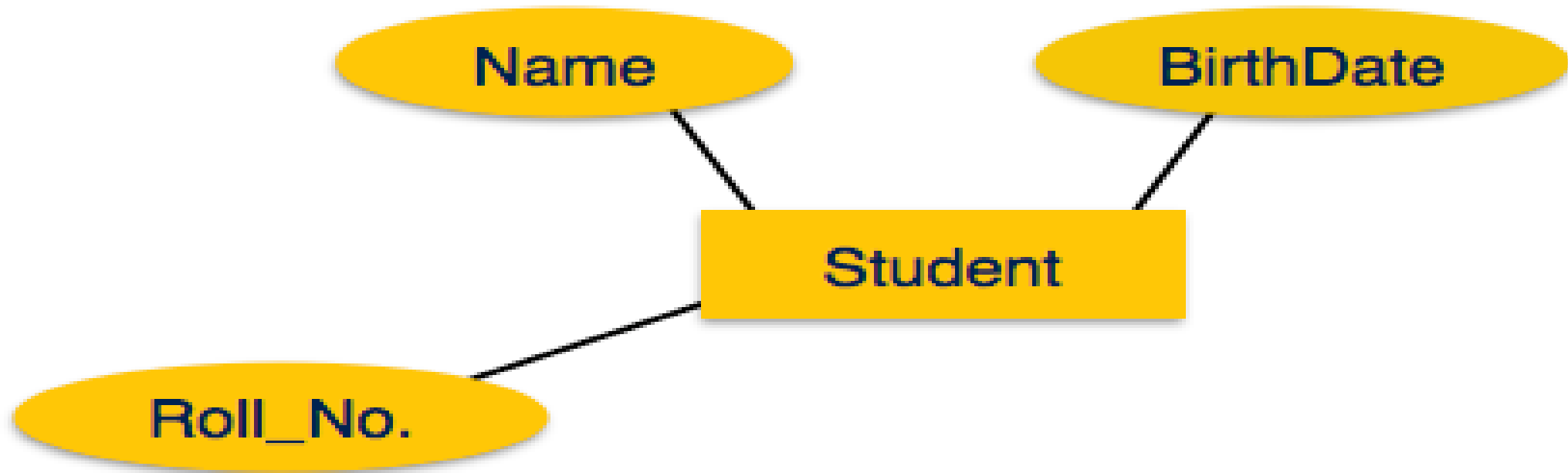
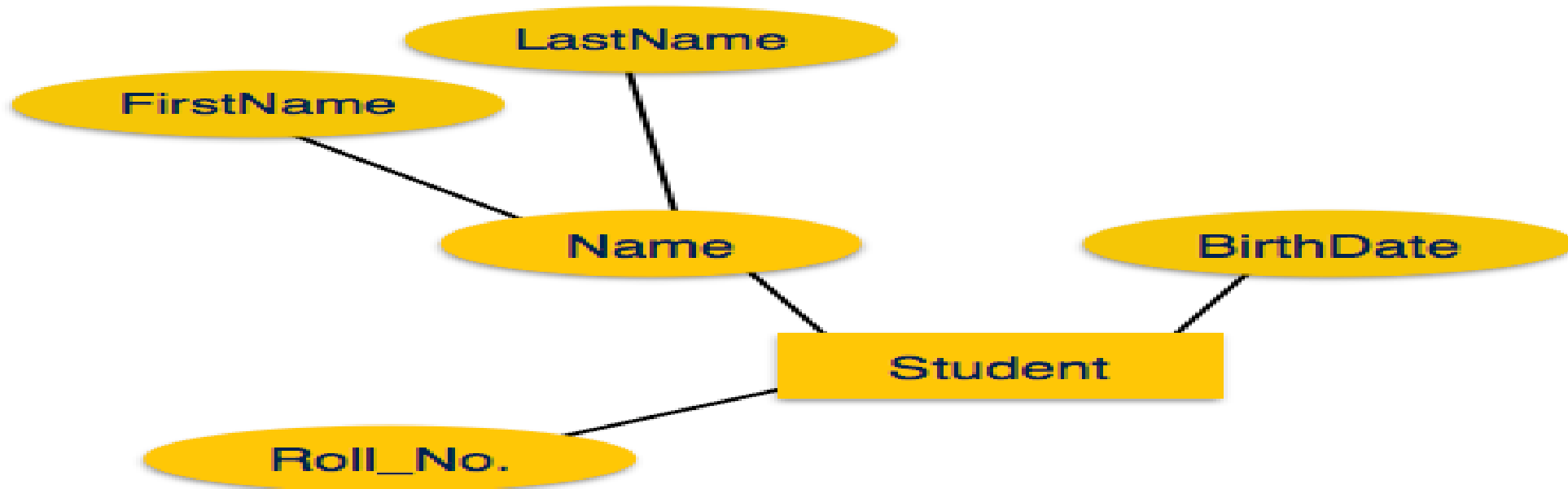# Attributes

- Attributes are the properties of entities. Attributes are represented by means of ellipses. Every ellipse represents one attribute and is directly connected to its entity (rectangle).
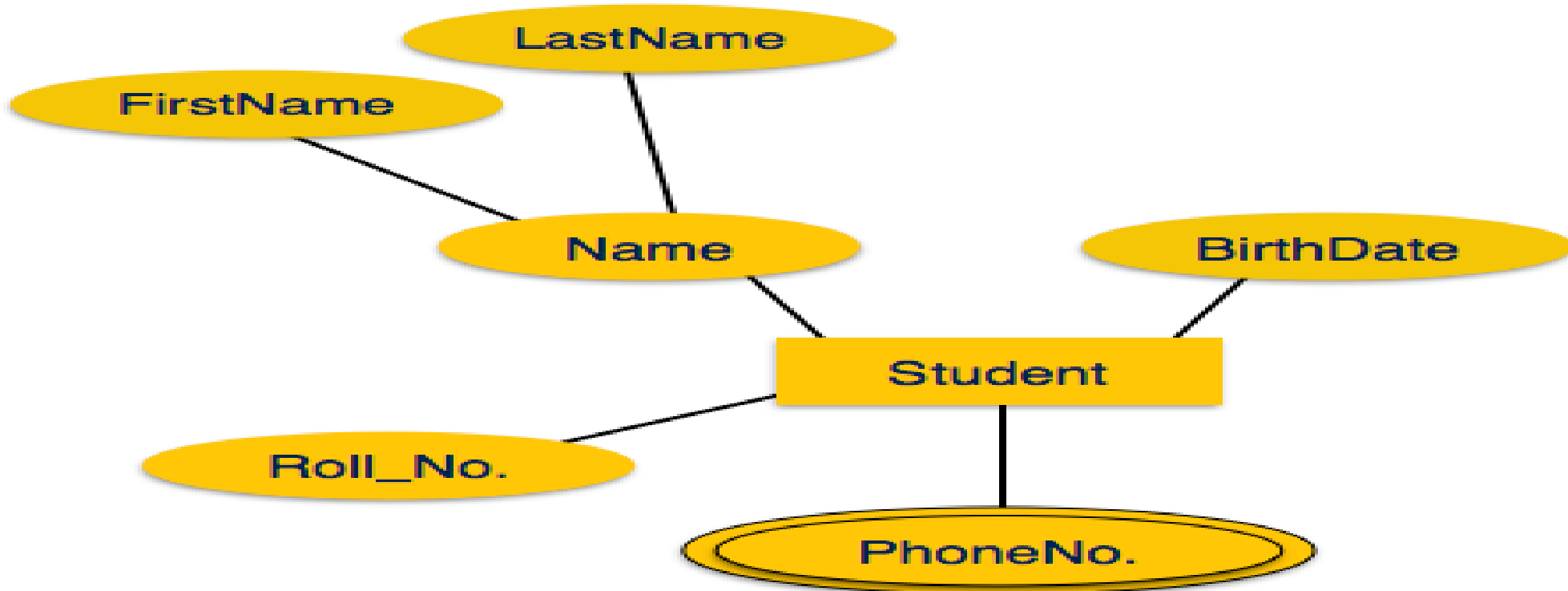
Attributes Cont..

- If the attributes are **composite**, they are further divided in a tree like structure.
- Every node is then connected to its attribute. That is, composite attributes are represented by ellipses that are connected with an ellipse.
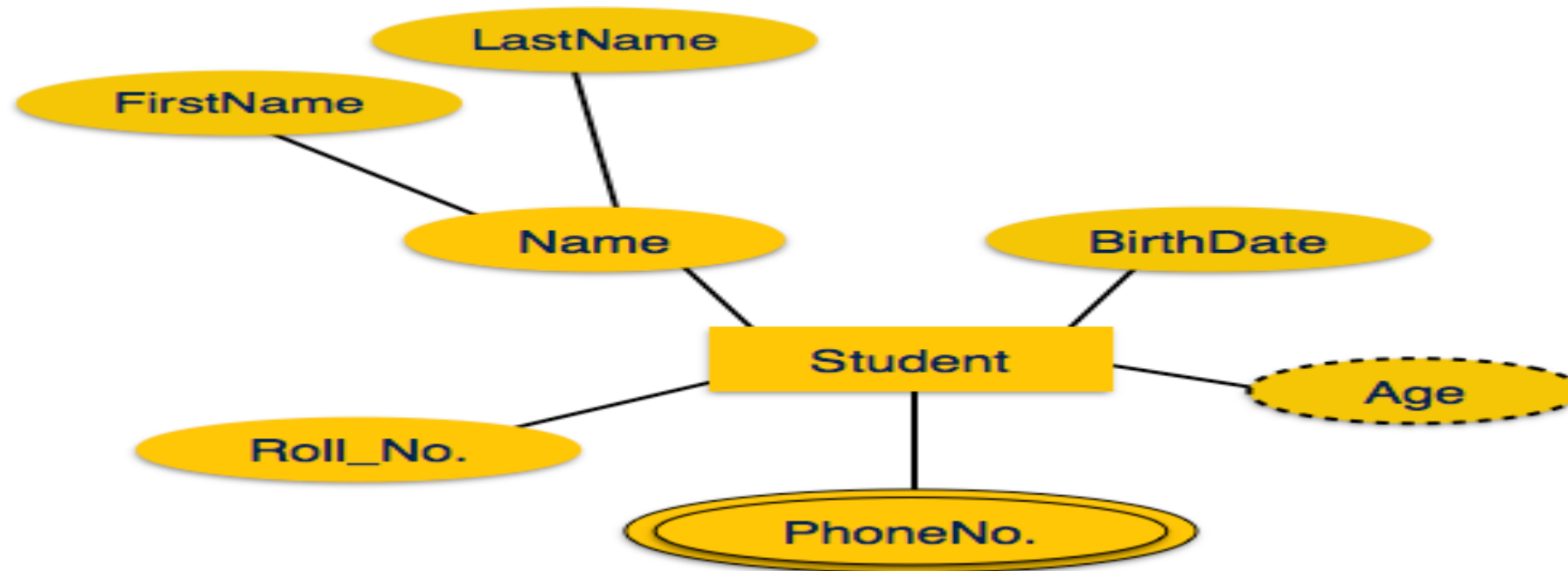
CSIT IIT UNIT- 9 Fundamentals of Database

3/25/2023

Attributes Cont..

- **Multivalued** attributes are depicted by double ellipse.

CSIT IIT UNIT- 9 Fundamentals of Database

3/25/2023

Attributes Cont..

- **Derived** attributes are depicted by dashed ellipse

Relationship

- Relationships are represented by diamond-shaped box. Name of the relationship is written inside the diamond-box.

- All the entities (rectangles) participating in a relationship, are connected to it by a line.

- Binary Relationship and Cardinality

- A relationship where two entities are participating is called a **binary relationship**.

- Cardinality is the number of instance of an entity from a relation that can be associated with the relation.

Relationship Cont...
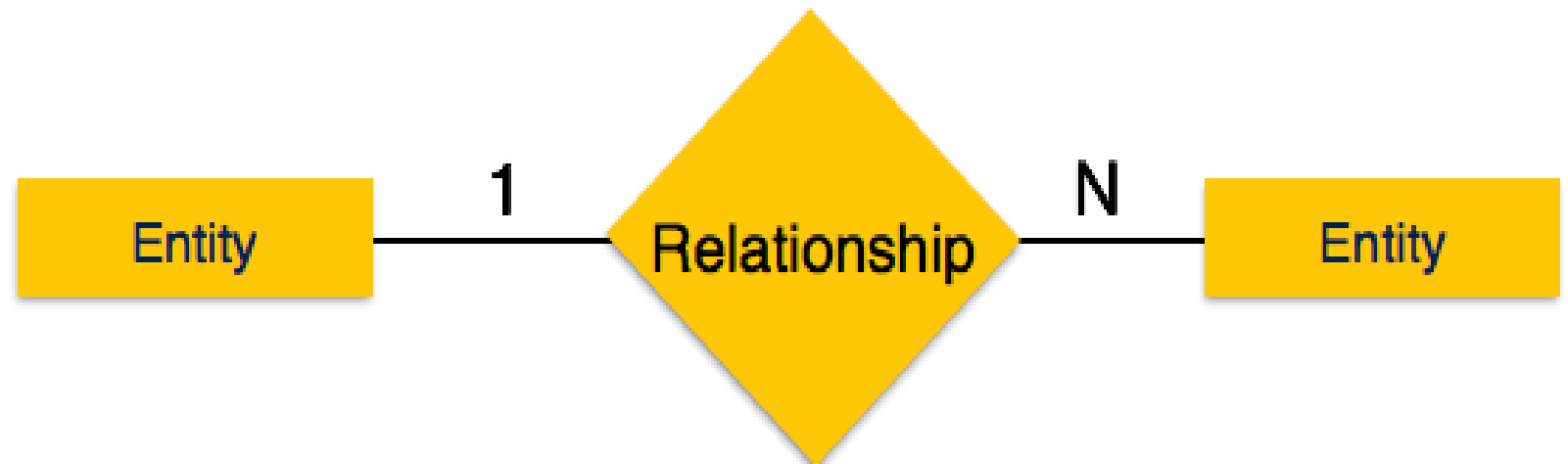
One-to-one − When only one instance of an entity is associated with the relationship, it is marked as '1:1'.

The following image reflects that only one instance of each entity should be associated with the relationship. It depicts one-to-one relationship.

CSIT IIT UNIT- 9 Fundamentals of Database

3/25/2023
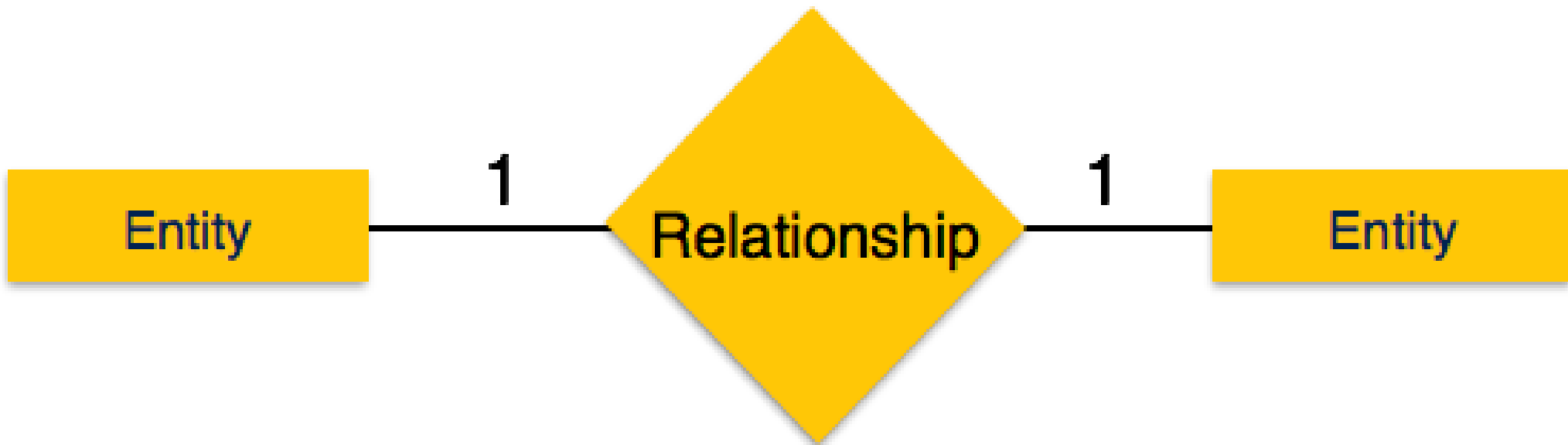
Relationship Cont…

- **One-to-many** – When more than one instance of an entity is associated with a relationship, it is marked as '1:N'.

- The following image reflects that only one instance of entity on the left and more than one instance of an entity on the right can be associated with the relationship. It depicts one-to-many relationship.

Relationship Cont…

- **Many-to-one** – When more than one instance of entity is associated with the relationship, it is marked as 'N:1'.

- The following image reflects that more than one instance of an entity on the left and only one instance of an entity on the right can be associated with the relationship. It depicts many-to-one relationship.

CSIT IIT UNIT- 9 Fundamentals of Database

3/25/2023
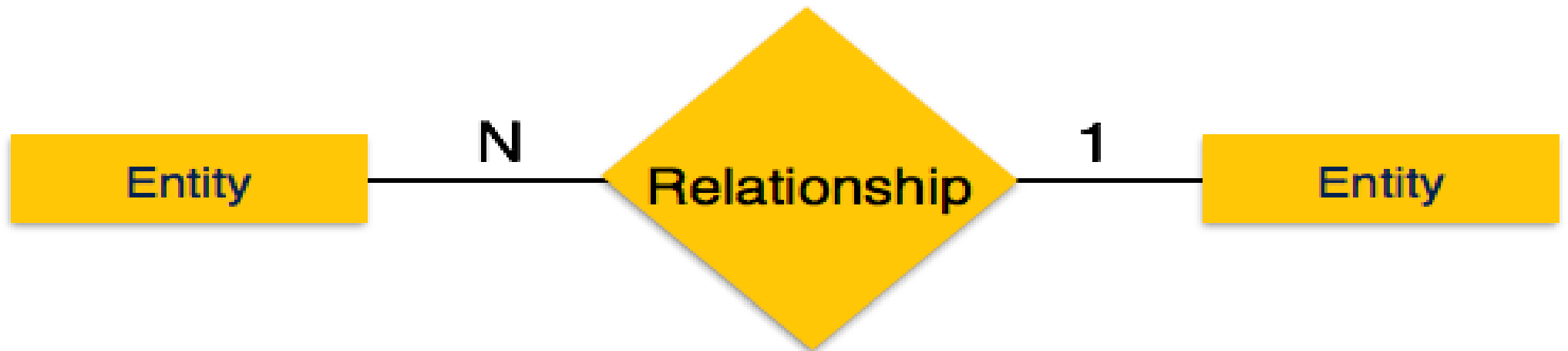
Relationship Cont…

- **Many-to-many** – The following image reflects that more than one instance of an entity on the left and more than one instance of an entity on the right can be associated with the relationship. It depicts many-to-many relationship.
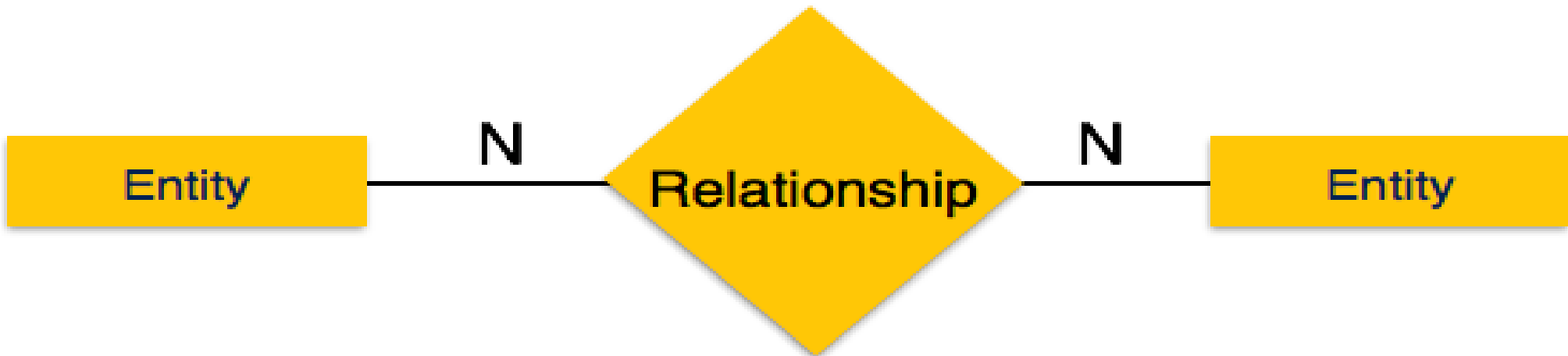
Database Models Cont..

2. Relational Model

- The most popular data model in DBMS is the Relational Model. It is more scientific model than others.

- This model is based on first-order predicate logic and defines a table as an n-ary relation

- **Relational Model** (RM) represents the database as a collection of relations. A **relation** is nothing but a table of values.

- Every row in the table represents a collection of related data values. These rows in the table denote a real-world entity or relationship. ... The data are represented as a set of relations

- In **relational model**, the **data** and relationships are represented by collection of inter-related tables.

- Each table is a group of column and rows, where column represents attribute of an entity and rows represents records

Roll_No.	Name	Department
101	Steive	Comp. Sci.
265	Jhoson	Finance
505	Margret	Biology
325	Jenny	Social Sci.
256	Davis	Comp. Sci.
453	Sheryl	Biology
365	Emma	Maths

Student Relation in Relational Model

Types of Database Management System

- Hierarchical Model
- Network Model
- Relational Model
- Object-oriented Model

Types of DBMS Cont..

Hierarchical DBMS

- In a Hierarchical database, model data is organized in a tree-like structure. Data is Stored Hierarchically (top down or bottom up) format.

- Data is represented using a parent-child relationship. In Hierarchical DBMS parent may have many children, but children have only one parent.

Network Model

- The network database model allows each child to have multiple parents. It helps you to address the need to model more complex relationships like as the orders/parts many-to-many relationship.

- In this model, entities are organized in a graph which can be accessed through several paths.

Types of DBMS Cont..

Relational model

- Relational DBMS is the most widely used DBMS model because it is one of the easiest.

- This model is based on normalizing data in the rows and columns of the tables. Relational model stored in fixed structures and manipulated using SQL.

Object-Oriented Model

- In Object-oriented Model data stored in the form of objects. The structure which is called classes which display data within it.

- It defines a database as a collection of objects which stores both data members values and operations.

Advantages of DBMS

- DBMS offers a variety of techniques to store & retrieve data
- DBMS serves as an efficient handler to balance the needs of multiple applications using the same data
- Uniform administration procedures for data
- Application programmers never exposed to details of data representation and storage.
- A DBMS uses various powerful functions to store and retrieve data efficiently.
- Offers Data Integrity and Security
- The DBMS implies integrity constraints to get a high level of protection against prohibited access to data.
- A DBMS schedules concurrent access to the data in such a manner that only one user can access the same data at a time
- Reduced Application Development Time

Disadvantage of DBMS

- Cost of Hardware and Software of a DBMS is quite high which increases the budget of your organization.

- Most database management systems are often complex systems, so the training for users to use the DBMS is required.

- In some organizations, all data is integrated into a single database which can be damaged because of electric failure or database is corrupted on the storage media

- Use of the same program at a time by many users sometimes lead to the loss of some data.

- DBMS can't perform sophisticated calculations

DBMS and it's services

- Transaction Processing:
- Concurrency Management
- Recovery
- Security
- Language Interface
- Data catalog
- Storage Management

Quality of Information

- Accuracy
- Timeliness
- Relevancy

Normalization

- **Normalization** is a database design technique that reduces data redundancy and eliminates undesirable characteristics like Insertion, Update and Deletion Anomalies.

- **Normalization** rules divides larger tables into smaller tables and links them using relationships.

Normalization

- **Normalization** is a database design technique that reduces data redundancy and eliminates undesirable characteristics like Insertion, Update and Deletion Anomalies. Normalization rules divides larger tables into smaller tables and links them using relationships.

- The purpose of Normalization in SQL is to eliminate redundant (repetitive) data and ensure data is stored logically.

Types of Normalization

- First normal form(1NF)

- Second normal form(2NF)

- Third normal form(3NF)

- Boyce & Codd normal form (BCNF)

Starting with normal form	Convert to normal form	Abbreviation	Use the rule
Un-normalized data	First	1NF	remove repeating groups of data within a single tuple into new tuples with only one data value for each attribute
First	Second	2NF	Extract non-key partial dependencies (items not fully functionally dependent on the key) into a separate relation
Second	Third	3NF	remove transitive dependencies into a separate relation
Third	Boyce-Codd	BCNF	remove overlapping keys by identifying a single primary key and holding other values in a separate relation
Third Or Boyce-Codd	Fourth	4NF	remove multi-valued dependencies by extracting independent data into a separate relation
Fourth	Fifth	5NF	remove join dependencies caused by interdependent data

Advantages of Normalization

- The following are the advantages of the normalization.
- • More efficient data structure.
  • Avoid redundant fields or columns.
  • More flexible data structure i.e. we should be able to add new rows and data values easily
  • Better understanding of data.
  • Ensures that distinct tables exist when necessary.
- • Easier to maintain data structure i.e. it is easy to perform operations and complex queries can be easily handled.
  • Minimizes data duplication.
  • Close modeling of real world entities, processes and their relationships.

Disadvantages of Normalization

- The following are disadvantages of normalization.
  - You cannot start building the database before you know what the user needs.
  - On Normalizing the relations to higher normal forms i.e. 4NF, 5NF the performance degrades.
  - It is very time consuming and difficult process in normalizing relations of higher degree.
  - Careless decomposition may leads to bad design of database which may leads to serious problems.

Database Design

- Database design is the organization of data according to a database model.
- The designer determines what data must be stored and how the data elements interrelate.
- With this information, they can begin to fit the data to the database model. Database management system manages the data accordingly

Database Design

Database Design Process

Real World System

Database Representation

CSIT IIT UNIT- 9 Fundamentals of Database

3/25/2023
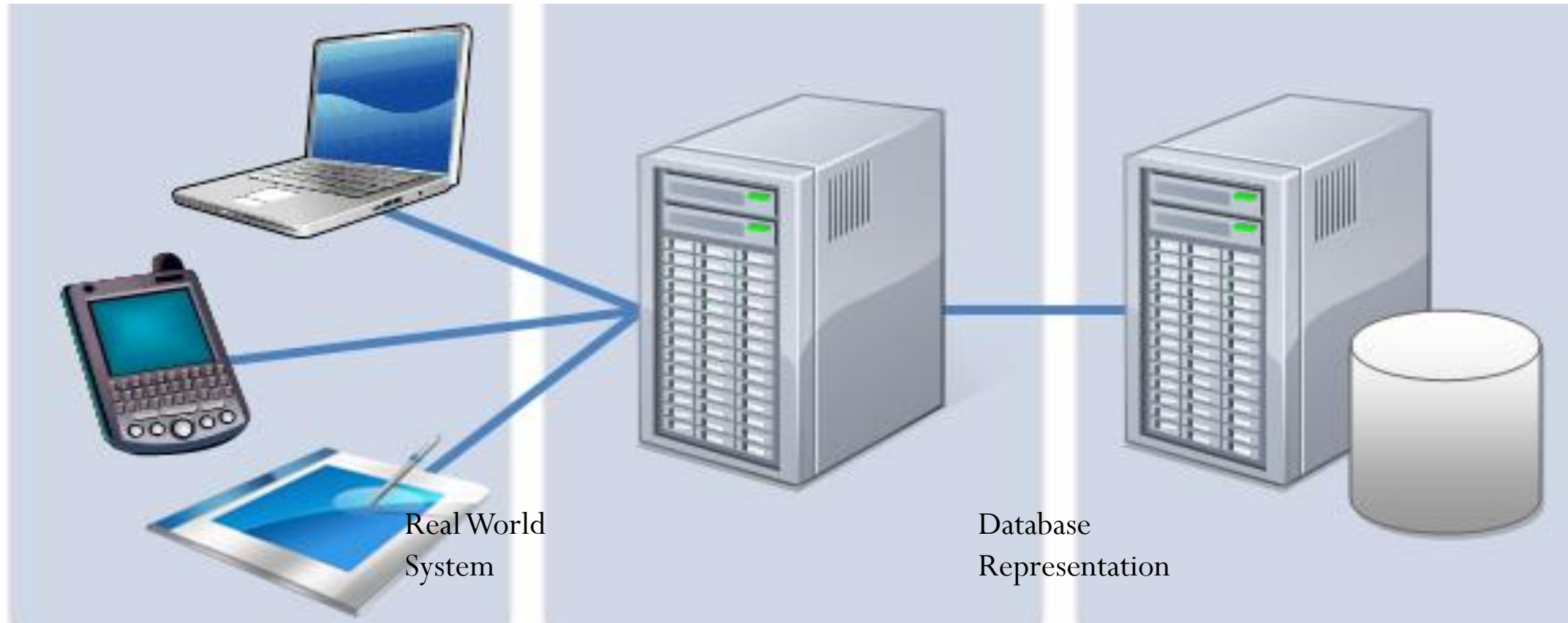
Database System Architecture

- Database architecture uses programming languages to design a particular type of software for businesses or organizations.

- Database architecture focuses on the design, development, implementation and maintenance of computer programs that store and organize information for businesses, agencies and institutions.

- A database architect develops and implements software to meet the needs of users.

Centralized Database

- A centralized database is basically a type of database that is stored, located as well as maintained at a single location only. This type of database is modified and managed from that location itself. This location is thus mainly any database system or a centralized computer system. The centralized location is accessed via an internet connection (LAN, WAN, etc). This centralized database is mainly used by institutions or organizations.

- This location is most often a central computer or database system, for example a desktop or server CPU, or a mainframe computer.

CSIT IIT UNIT- 9 Fundamentals of Database
3/25/2023
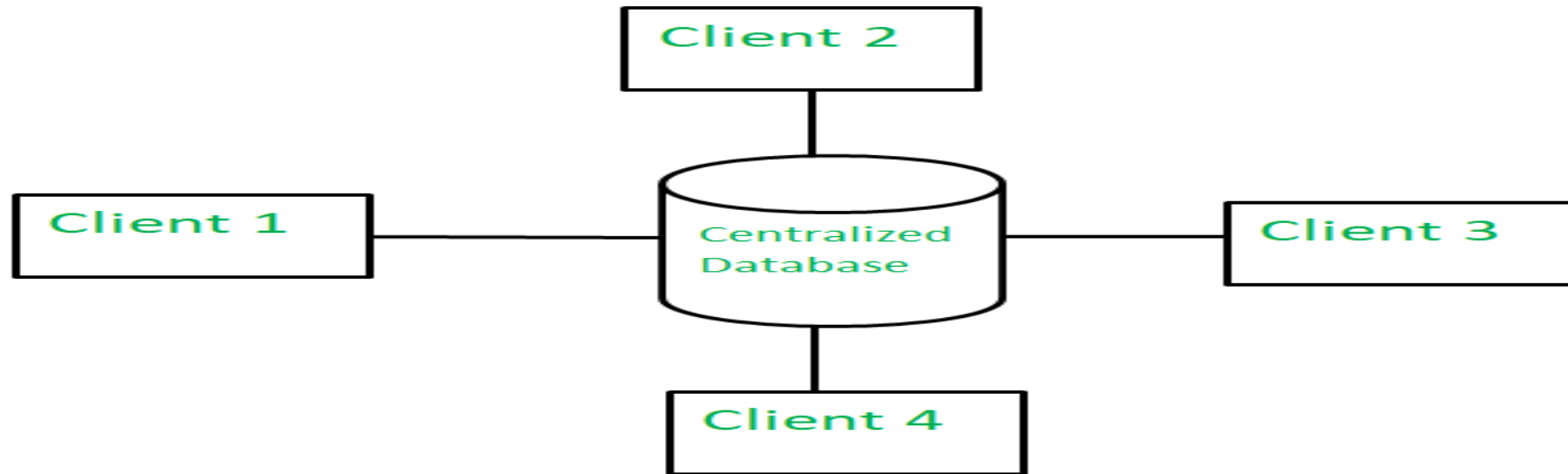
Centralized Database Cont..

Advantages –

- Since all data is stored at a single location only thus it is easier to access and co-ordinate data.
- The centralized database has very minimal data redundancy since all data is stored at a single place.
- It is cheaper in comparison to all other databases available.

Disadvantages –

- The data traffic in case of centralized database is more.
- If any kind of system failure occurs at centralized system then entire data will be destroyed.

Client /Server Database

- A client-server database is **one where the database resides on a server**, and client applications are written to access the database.

- **Client/server** is developed to deal with various computing environments that have a large number of computers and servers connected together via a network.

- The **client-server** model describes how a **server** provides resources and services to one or more **clients**.

- **Examples** of servers include web servers, mail servers, and file servers. Each of these servers provide resources to **client** devices, such as desktop computers, laptops, tablets, and smartphones

Client /Server Database
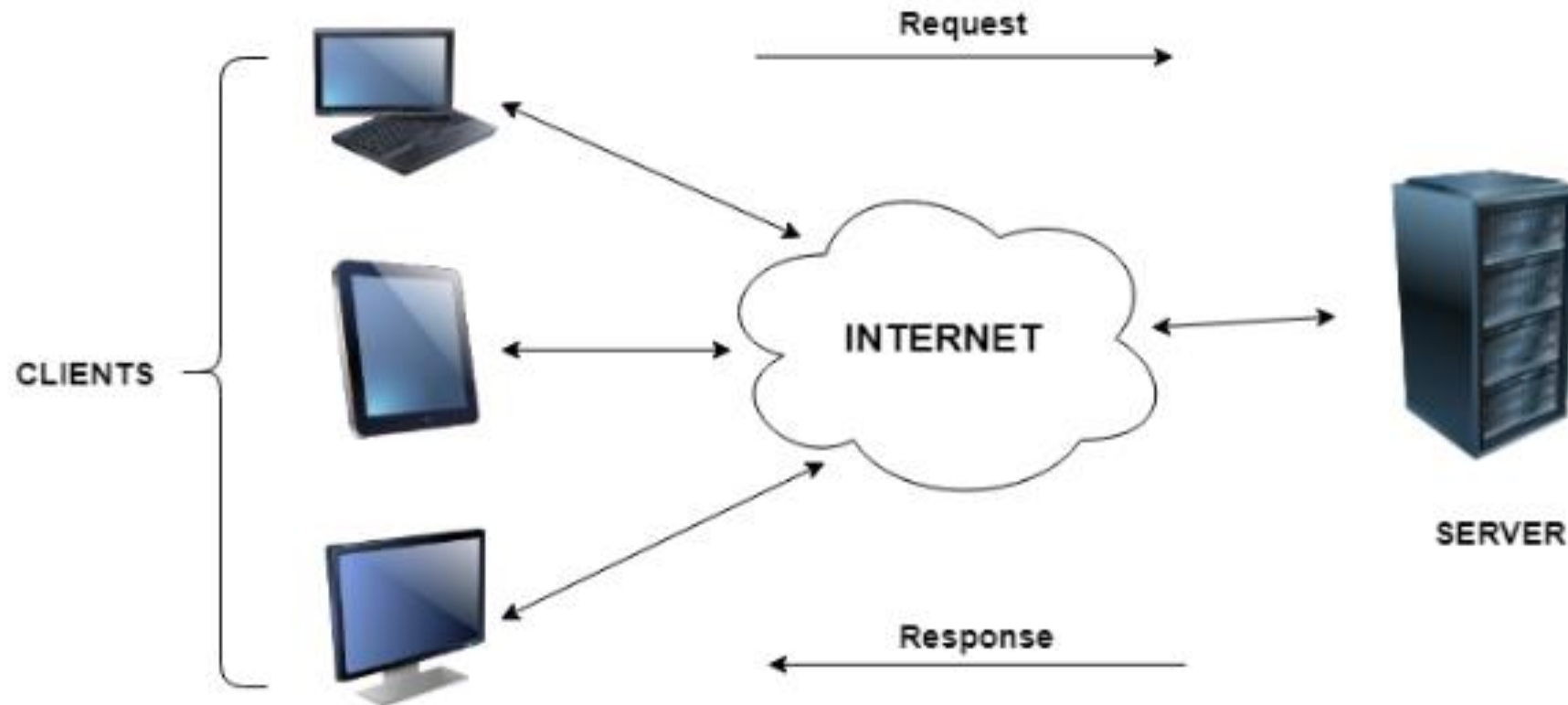
Client /Server Database

Advantages

- All the required data is concentrated in a single place i.e. the server. So it is easy to protect the data and provide authorization and authentication.
- The server need not be located physically close to the clients. Yet the data can be accessed efficiently.
- It is easy to replace, upgrade or relocate the nodes in the client server model because all the nodes are independent and request data only from the server.
- All the nodes i.e clients and server may not be build on similar platforms yet they can easily facilitate the transfer of data.

Disadvantages

- If all the clients simultaneously request data from the server, it may get overloaded. This may lead to congestion in the network.
- If the server fails for any reason, then none of the requests of the clients can be fulfilled. This leads of failure of the client server network.
- The cost of setting and maintaining a client server model are quite high

Distributed Database :

- A distributed database is basically a type of database which consists of multiple databases that are connected with each other and are spread across different physical locations.

- The data that is stored on various physical locations can thus be managed independently of other physical locations.

- The communication between databases at different physical locations is thus done by a computer network.
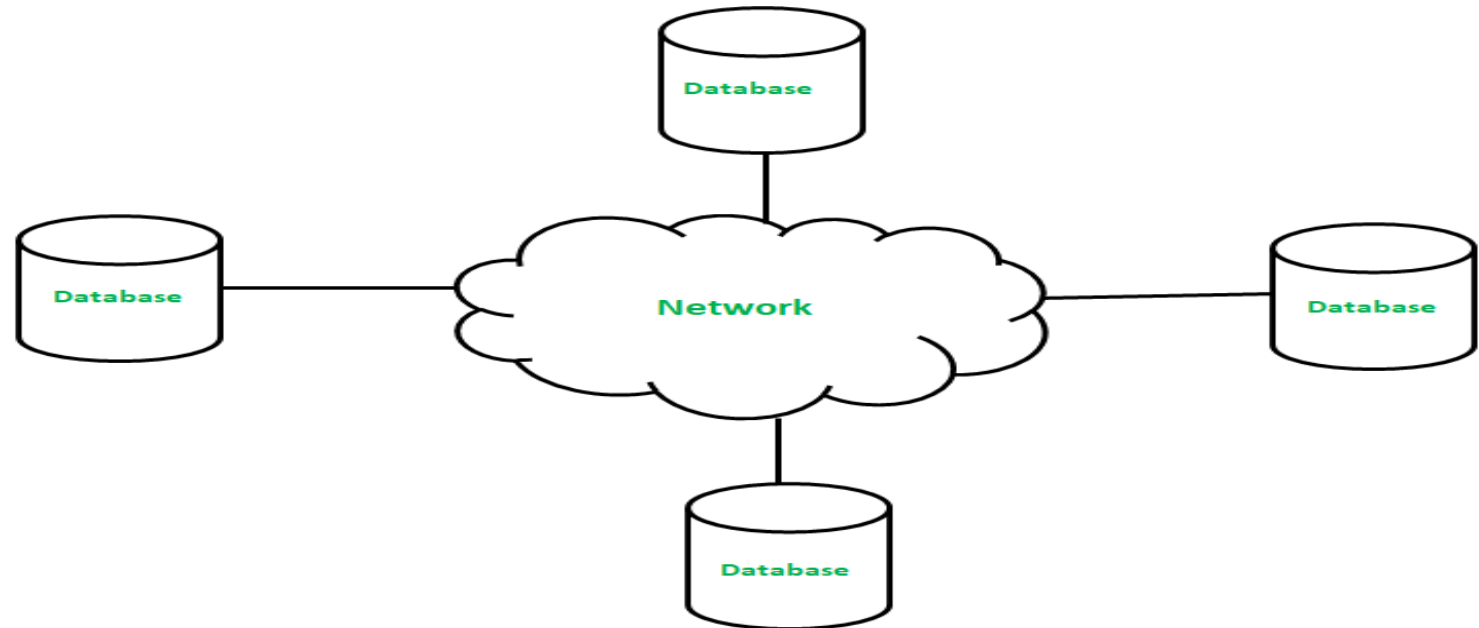
Distributed Database :

Advantages –

- This database can be easily expanded as data is already spread across different physical locations.

- The distributed database can easily be accessed from different networks.

- This database is more secure in comparison to centralized database.

Disadvantages –

- This database is very costly and it is difficult to maintain because of its complexity.

- In this database, it is difficult to provide a uniform view to user since it is spread across different physical locations.

Database Security

- Database security involves protecting a database from unauthorized access, malicious destruction and even any accidental loss or misuse.

- Due to the high value of data incorporate databases, there is strong motivation for unauthorized users to gain access to it, for instance, competitors or dissatisfied employees

Database Security Cont…

There are various ways how we can secure our system.

1. Authorization

- Authorization is the granting of a right or privilege that enables a subject to have legitimate access to a system or a system's object.

Usually, a user or subject can gain access to or a system through individual user accounts where each user is given a unique identifier, which is used by the operating system to determine that they have the authorization to do so.

2. Access Control

- Access controls to a database system is based on the granting and revoking of privileges. A privilege allows a user to create or access (that is read, write or modify) a database object or to execute a DBMS utility.

- The DBMS keeps track of how these privileges are granted to users and possibly revoked, and ensures that at all times only users with necessary privileges can access an object.

Database Security Cont…

3. Views

- A view is created by querying one or more of the base tables, producing a dynamic result table for the user at the time of the request.

- The user may be allowed to access the view but not the base tables which the view is based.

- The view mechanism hides some parts of the database from certain users and the user is not aware of the existence of any attributes or rows that are missing from the view

4. Backup and recovery

- Backup is the process of periodically taking a copy of the database and log file to offline storage media.

- Backup is very important for a DBMS to recover the database following a failure or damage

Database Security Cont…

5. Encryption

- Encryption is the process of encoding of the data using a special algorithm that renders the data unreadable by any program without the decryption key .

- Data encryption can be used to protect highly sensitive data like customer credit card numbers or user password.

- Some DBMS products include encryption routines that would automatically encode the sensitive data when they are stored or transmitted over communication channels

## 6. RAID (Redundant Array of Independent Disks)

- The DBMS should continue to operate even though if one of the hardware components fails.

- The hardware that the DBMS is running on must be fault-tolerant where the DBMS should continue operating and processing even if there is hardware failure.

- The main hardware components that should be fault-tolerant are disk drives, disk controllers, CPU, power supplies and cooling fans
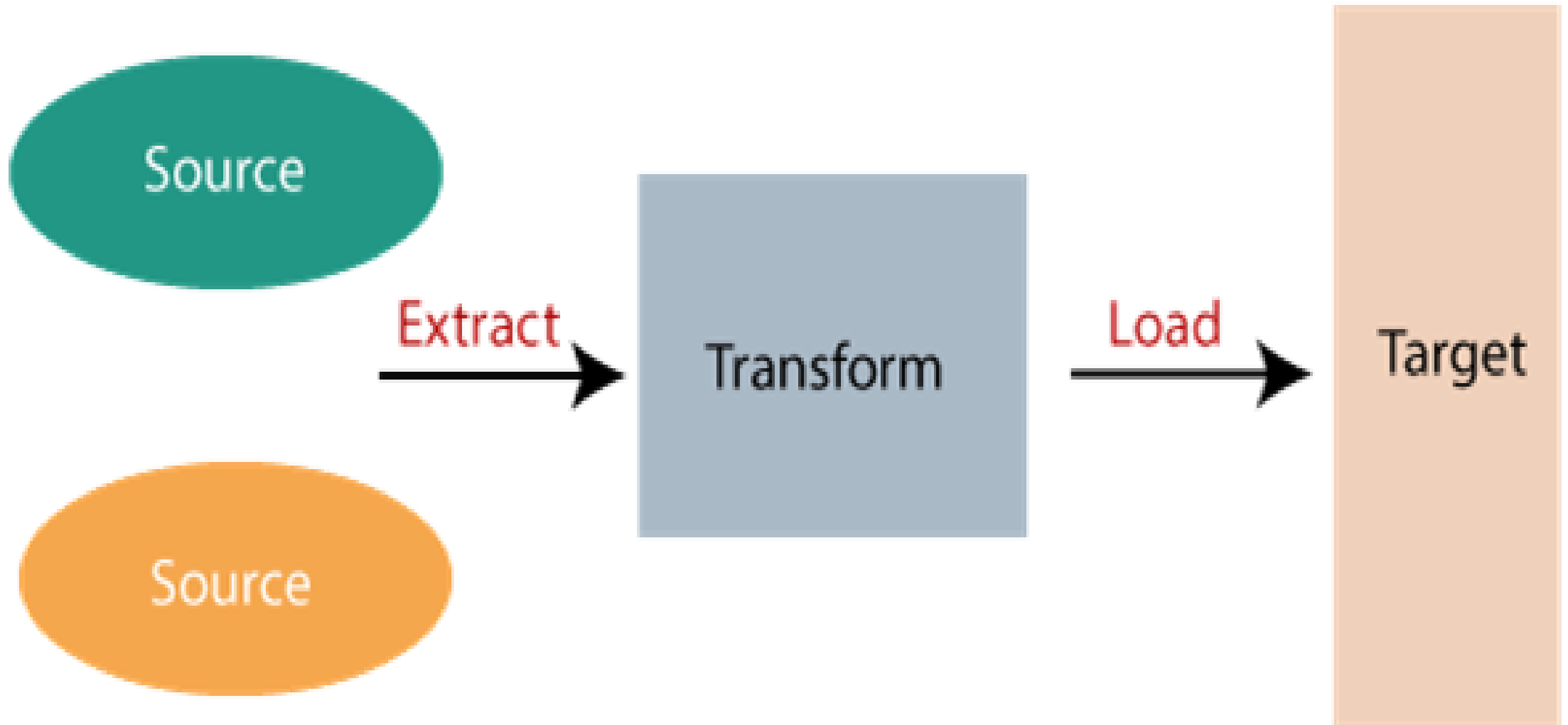
Data Warehouse

- A data warehouse is a large repository of data which comes from operational sources and has four properties:
  - Non volatile
  - Time varying
  - Subject oriented
  - Integrated

Data Warehouse and Data Mining

- A Data warehouse is an information system that contains historical and commutative data from single or multiple sources.

- Data Warehouse Concepts simplify the reporting and analysis process of organizations.
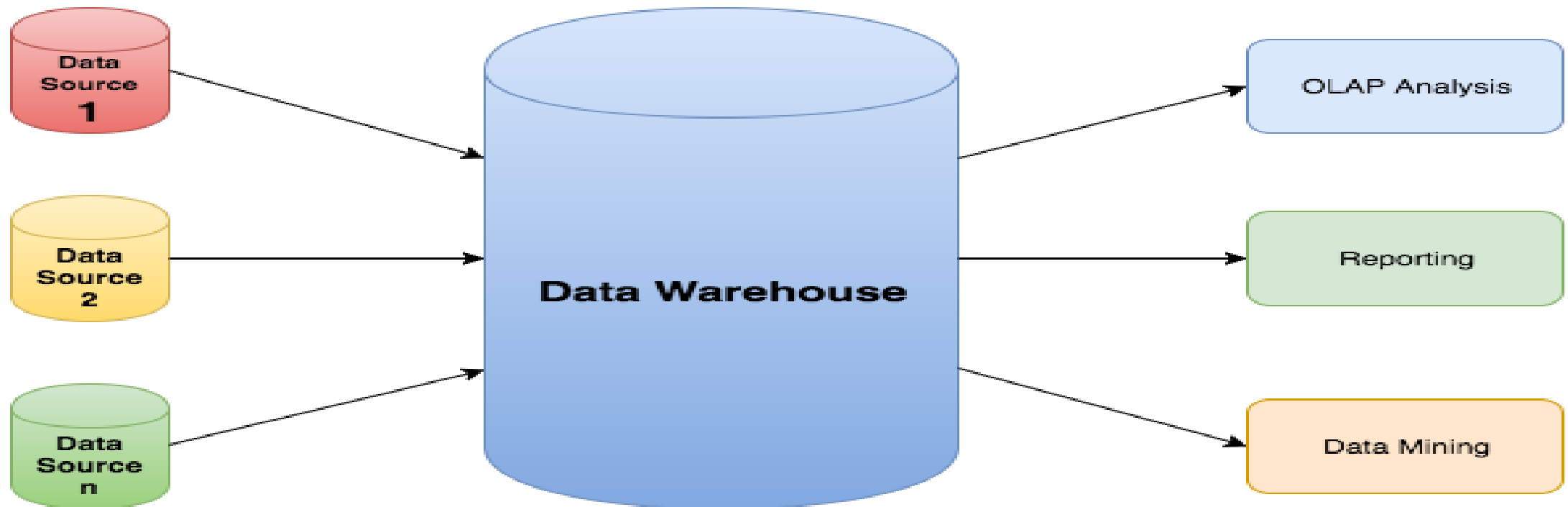
Data Warehousing Process

Data Warehouse Cont...

- The collection of data used by data warehouse may be characterized as subject-oriented, integrated, non volatile and time-variants.

1. Subject Oriented

- Data is arranged and optimized to provide answer to questions from diverse functional areas. Data is organized and summarized by topic like Sales/Marketing/Finance/Distribution etc

2. Integrated

- The data warehouse is a centralized, consolidated database that integrates data derived from the entire organization.
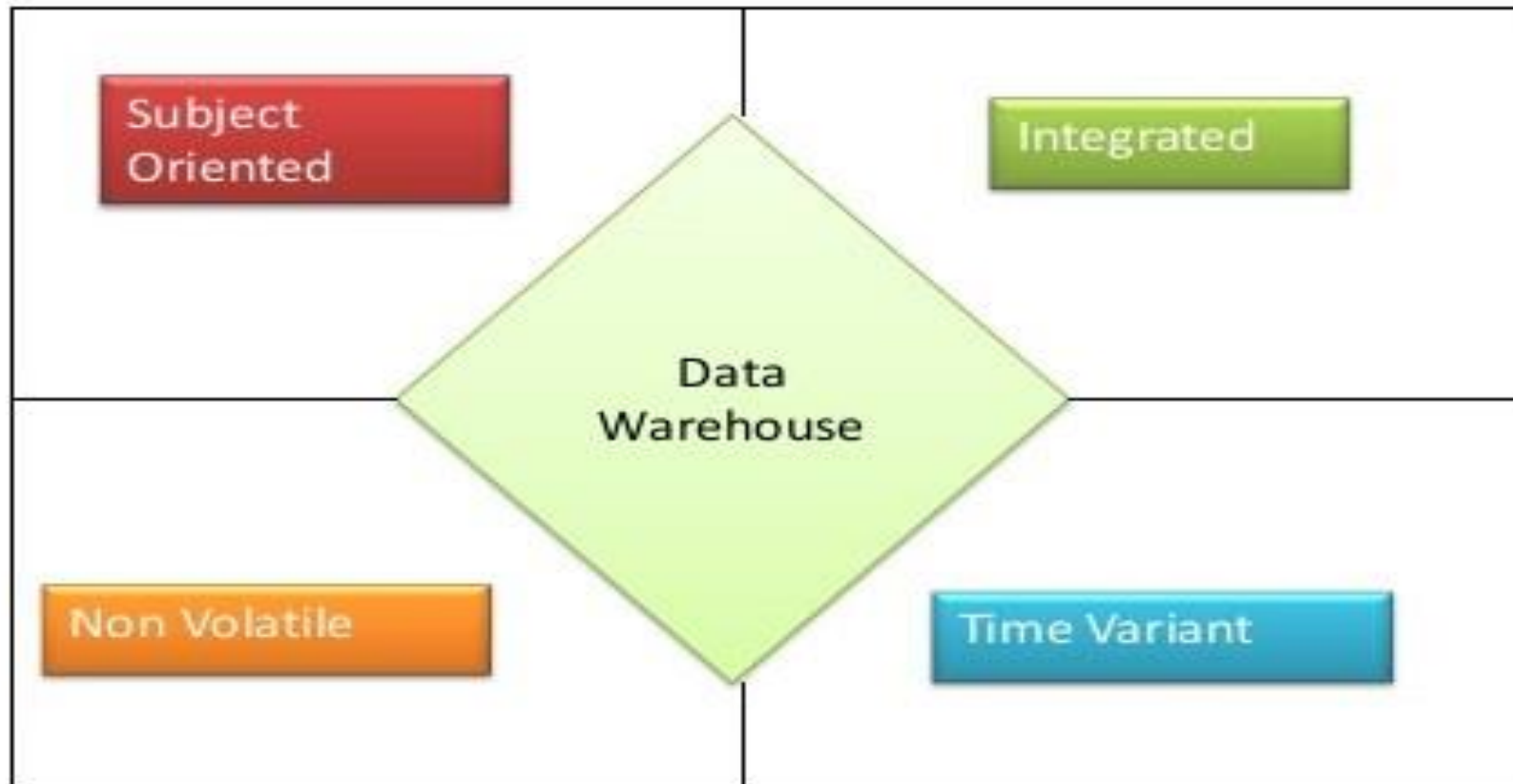
Data Warehouse Cont…

4. Time Variant

- The Data Warehouse represents the flow of data through time. It contains projected data from statistical models. Data is periodically uploaded then time-dependent data is recomputed.

5. Non-Volatile

- Once data is entered it is NEVER removed. It represents the company's entire history—Near term history is continually added to it. It is always growing and must support terabyte databases and multiprocessors. It is Read-Only database for data analysis and query processing

Properties of Data warehouse

CSIT IIT UNIT- 9 Fundamentals of Database
3/25/2023

Data Warehouse Architecture

- A **data warehouse architecture** is a method of defining the overall **architecture** of **data** communication processing and presentation that exist for end-clients computing within the enterprise.

- Each **data warehouse** is different, but all are characterized by standard vital components.

- The main benefits of implementing a data warehouse are cost-effective decision-making, better business intelligence, enhanced customer service, business re-engineering, information system re-engineering etc

Architecture of a Data Warehouse
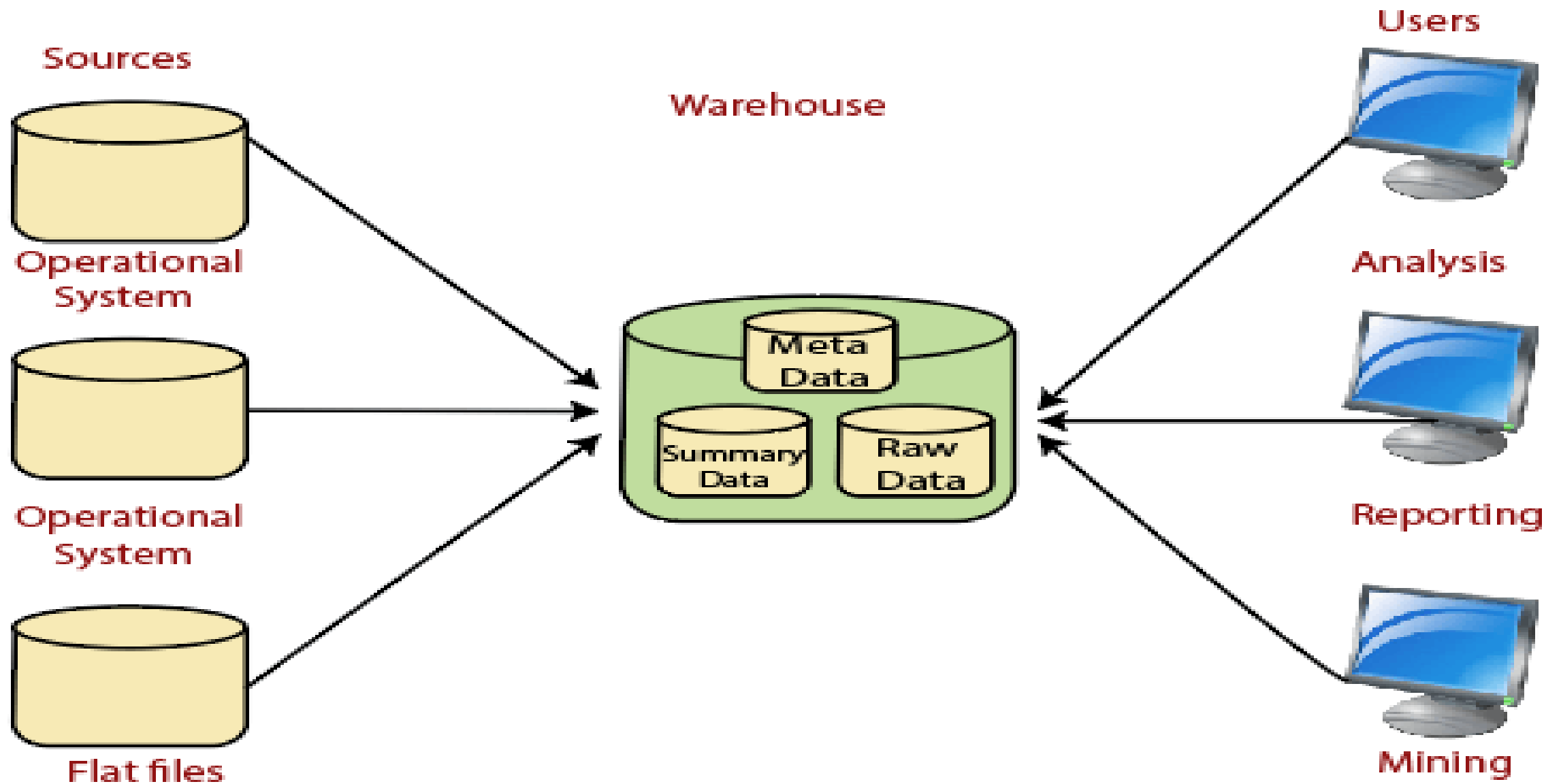
Advantages and Disadvantages of a Data Warehouse

Advantages

- Enables Historical Insight
- Enhances Conformity and Quality of Data
- Boosts Efficiency
- Increase the Power and Speed of Data Analytics.
- Drives Revenue.
- Scalability.
- Interoperates with On-Premise and Cloud.
- Data Security.

Disadvantages

- Query-driven approach needs complex integration and filtering processes.
- This approach is very inefficient.
- It is very expensive for frequent queries.
- This approach is also very expensive for queries that require aggregations.
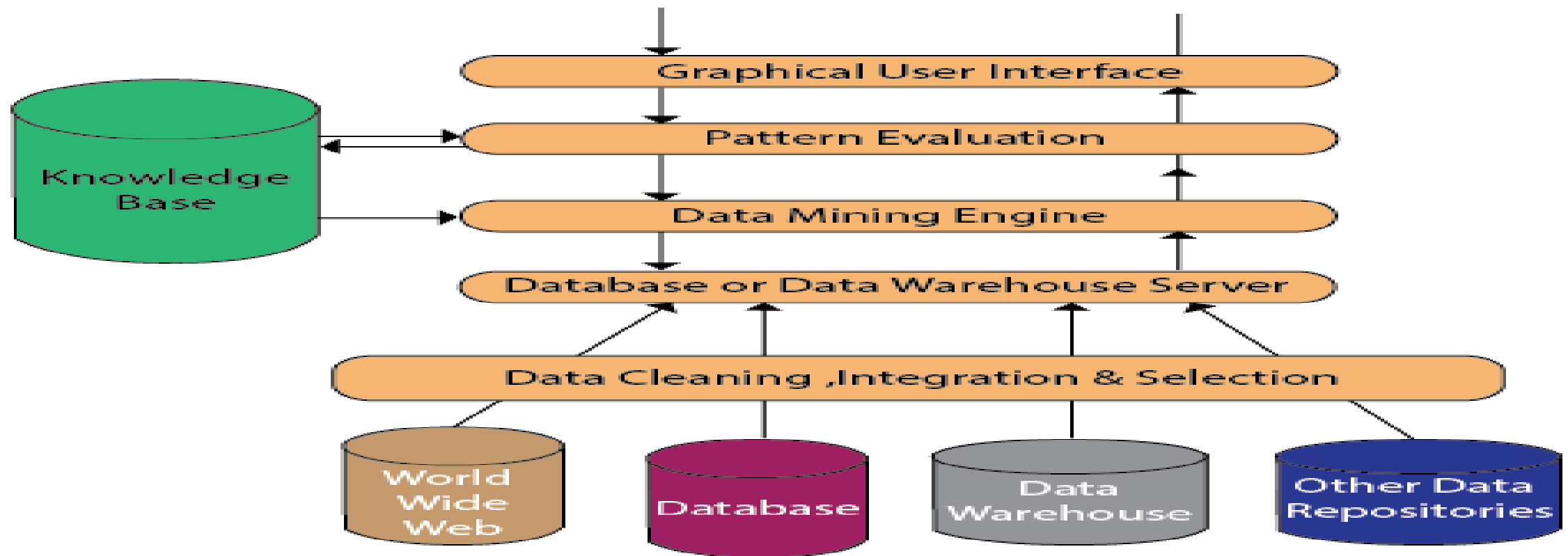
Data Mining

- It is a process of semi automatically finding of useful information from large repositories.
- It refers to extracting or "mining" knowledge from large amounts of data.
- Interesting knowledge, regularities can be extracted from databases and viewed from different angles.
- Applied to decision making, process control, information management, and query processing
- Data mining is a step in the Knowledge Discovery Process

CSIT IIT UNIT- 9 Fundamentals of Database
3/25/2023

Data Mining Cont..

- Data mining refers to extracting or mining knowledge from large amounts of data.
- It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.
- The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.
- The key properties of data mining are
    1. Automatic discovery of patterns
    2. Prediction of likely outcomes
    3. Creation of actionable information
    4. Focus on large datasets and databases

Architecture of Data Mining

- The significant components of data mining systems are a data source, data mining engine, data warehouse server, the pattern evaluation module, graphical user interface, and knowledge base.

Advantages and Disadvantages of Data Mining Techniques

Advantages:

- It is helpful to predict future trends
- It signifies customer habits:
- Helps in decision making
- Increase company revenue
- It depends upon market-based analysis
- Quick fraud detection

Disadvantages:

- Cost. **Data mining** involves lots of technology in use for the **data** collection process. ...
- Security. Identity theft is a big issue when using **data mining**. ...
- Privacy. When using **data mining** there are many privacy concerns raised. ...
- Accuracy. ...
- Technical Skills. ...
- Information Misuse. ...
- Additional Information.

Big Data

- Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software. **Big Data** meaning a **data** that is huge in size.

- **Big data** is a term used to describe a collection of **data** that is huge in size and yet growing exponentially with time. **Big Data** analytics **examples** includes stock exchanges, social media sites, jet engines, etc

Database Administrator

- A database administrator (DBA) is a specialized computer systems administrator who maintains a successful database environment by directing or performing all related activities to keep the data secure.

- The top responsibility of a DBA is to maintain data integrity.

- This means the DBA will ensure that data is secure from unauthorized access but is available to users.

- DBA is responsible for backing up systems in case of power outages or other disasters.

- A DBA is also frequently involved in tasks related to training employees in database management and use, designing, implementing, and maintaining the database system and establishing policies and procedures related to the organization's data management policy.

Database Administrator Cont..

- Database administrator can be classified as

1. System DBA

- Overview System DBAs typically have a background in system architecture and are responsible for the physical and technical aspects of a database.

- This can include installing upgrades and patches to fix program bugs and ensuring that the database works properly in a firm's computer system.

2. Application DBA

- Overview Application DBAs use complex programming languages to write or debug programs that work with the database.

- Usually this database has been designed for a specific application or a set of applications, such as customer service software

Functions of DBA

- Software installation and Maintenance. A DBA often collaborates on the initial installation and configuration of a new Oracle, SQL Server etc database.

- Data Extraction, Transformation, and Loading.

- Specialized Data Handling.

- Database Backup and Recovery.

- Security.

- Authentication.

- Capacity Planning.

- Performance Monitoring.

CSIT IIT UNIT- 9 Fundamentals of Database