Explain Join Dependency with example.

Join Dependency means re-creating the original Table by joining multiple sub-tables of the given Table. It is a further generalization of MVD(multi-valued Dependencies). Join dependency is used for the loss-less decomposition of the Data.

Mathematical Representation:

R=(R1 ⋈ R2 ⋈ R3 ⋈ ………Rn) where R1,R2,R3…..Rn are sub-relation of R and ⋈ is **Natural Join** Operator.

Consider a table R as given below. It has 3 attributes I. e. X, Y, and Z as shown below. Decompose the given table into R1 which has 2 columns i.e. X and Y and R2 which has 2 columns i.e. Y and Z. Then combine R1 and R2 tables using Natural Join.

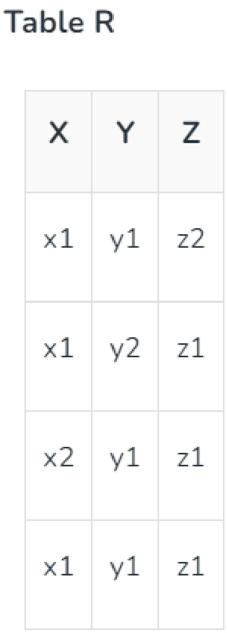


Table R

X	Y	Z
x1	y1	z2
x1	y2	z1
x2	y1	z1
x1	y1	z1

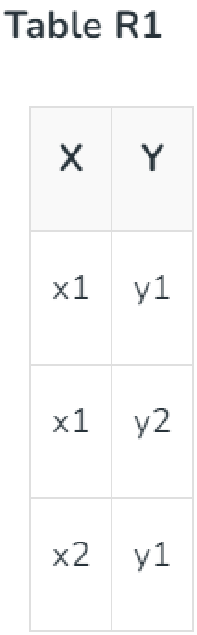


Table R1

X	Y
x1	y1
x1	y2
x2	y1

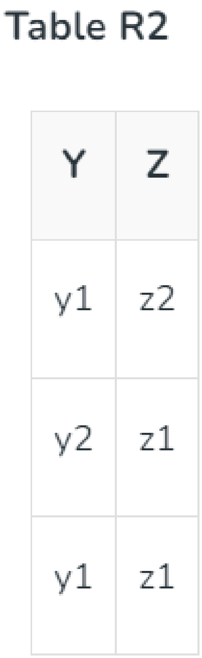


Table R2

Y	Z
y1	z2
y2	z1
y1	z1

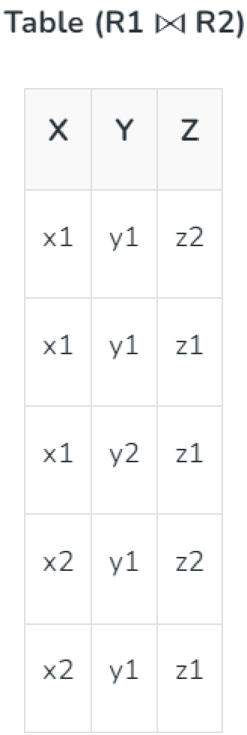


Table (R1 ⋈ R2)

X	Y	Z
x1	y1	z2
x1	y1	z1
x1	y2	z1
x2	y1	z2
x2	y1	z1

After combining both R1 and R2 we got unnecessary tuples which are also known as **Spurious Tuple.** To avoid this type of spurious tuple, we have to create another table R3 which will include columns X and Z. Now apply Natural Join on tables R1 , R2, and R3.

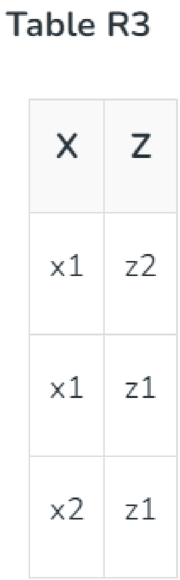


Table R3

X	Z
x1	z2
x1	z1
x2	z1

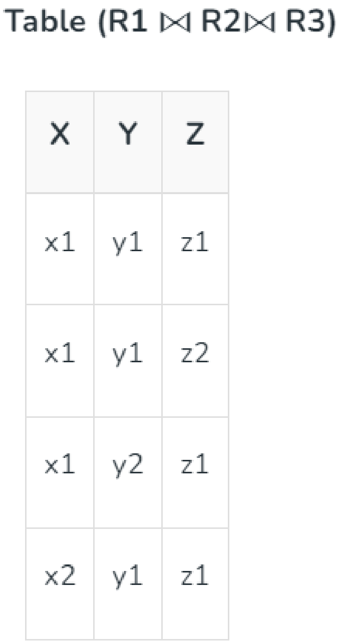


Table (R1 ⋈ R2⋈ R3)

X	Y	Z
x1	y1	z1
x1	y1	z2
x1	y2	z1
x2	y1	z1

The original table and the above-obtained table have the same data. In the above example, **(R1⋈R2⋈R3)=R holds** true.

Explain the Limitations of Relational Data Model.

1. **Complex Design:** Designing a relational database schema can be complex, especially for large and intricate applications.
2. **Fixed Schema:** They have a fixed schema, meaning the structure of the database is defined beforehand. Adding new columns or altering the schema often requires modifying existing applications, which can be cumbersome and may lead to downtime during updates.
3. **Performance Bottlenecks:** Certain operations, such as complex joins, can lead to performance bottlenecks in RDBMS, especially when dealing with large datasets.
4. **High Overhead:** They typically have more overhead compared to some NoSQL databases.
5. **Cost:** Some commercial RDBMS solutions can be expensive, especially when considering licensing, maintenance, and hardware requirements.
6. **Data Type Limitations:** RDBMS have predefined data types, and accommodating certain data formats or unstructured data can be difficult.
7. **Concurrency Issues:** In heavily concurrent systems, managing locks and ensuring data consistency can be complex and may lead to performance bottlenecks.
8. **Learning Curve:** They require users to be familiar with complex SQL queries and relational algebra.

Discuss the advantages of Query optimization.

- **Improved Query Performance:** By selecting the most efficient query plan, the RDBMS can execute the query in the shortest possible time.
- **Cost Reduction:** It reduces the cost of query execution by minimizing the use of system resources, such as CPU and memory

- **Query Tuning:** It provides a way to fine-tune queries to make them more efficient, even in complex and large-scale databases.
- **Reduced Complexity:** It makes the query execution process more efficient and reduces the complexity of the system.

How is Data Warehousing different from Data Base?

Database System	**Data Warehouse**
It supports operational processes.	It supports analysis and performance reporting.
Capture and maintain the data.	Explore the data.

Database System	Data Warehouse
100 MB to GB.	100 GB to TB.
ER based.	Star/Snowflake.
Application oriented.	Subject oriented.
Primitive and highly detailed.	Summarized and consolidated.
Flat relational.	Multidimensional.

Explain various normal forms with example.

1NF: A relation is in first normal form if every attribute in that relation is **singled valued attribute**.

Example: Consider following functional dependencies in relation R (A, B , C, D)

AB -> C
BC -> D

2NF: A relation is in second normal form if it is in 1NF and does not contain any partial dependency.

Example: Consider following functional dependencies in relation R (A, B , C, D)

AB -> C
BC -> D

3NF: A relation is said to be in third normal form, if it is in 2NF and does not have any transitive dependency for non-prime attributes.

Example: Consider relation R(A, B, C, D, E)

A -> BC,
CD -> E,
B -> D,
E -> A

BCNF: Conditions for BCNF

1. Table must be in 3NF.
2. In relation X->Y, X must be a superkey in a relation.

Example: Consider a relation R(A, B, C, D, E):

BC->D,

```
AC->BE,

B->E
```

4NF: Conditions for 4NF:

1. It must be in BCNF.
2. It does not have any multi-valued dependency.

STU_ID	COURSE	HOBBY
21	Computer	Dancing
21	Math	Singing
34	Chemistry	Dancing
74	Biology	Cricket
59	Physics	Hockey

To make the above table into 4NF, we can decompose it into two tables:

STUDENT_COURSE

STU_ID	COURSE
21	Computer
21	Math
34	Chemistry
74	Biology
59	Physics

STUDENT_HOBBY

STU_ID	HOBBY
21	Dancing
21	Singing
34	Dancing
74	Cricket
59	Hockey

5NF (Projected Normal Form): Conditions for 5NF:

1. It must be in 4NF.
2. It does not contain any join dependency and joining should be lossless.

SUBJECT	LECTURER	SEMESTER
Computer	Anshika	Semester 1
Computer	John	Semester 1
Math	John	Semester 1
Math	Akash	Semester 2
Chemistry	Praveen	Semester 1

To make the above table into 5NF, we can decompose it into three relations P1, P2 & P3:

P1

SEMESTER	SUBJECT
Semester 1	Computer
Semester 1	Math
Semester 1	Chemistry
Semester 2	Math

P2

SUBJECT	LECTURER
Computer	Anshika
Computer	John
Math	John
Math	Akash
Chemistry	Praveen

P3

SEMSTER	LECTURER
Semester 1	Anshika
Semester 1	John
Semester 1	John
Semester 2	Akash
Semester 1	Praveen

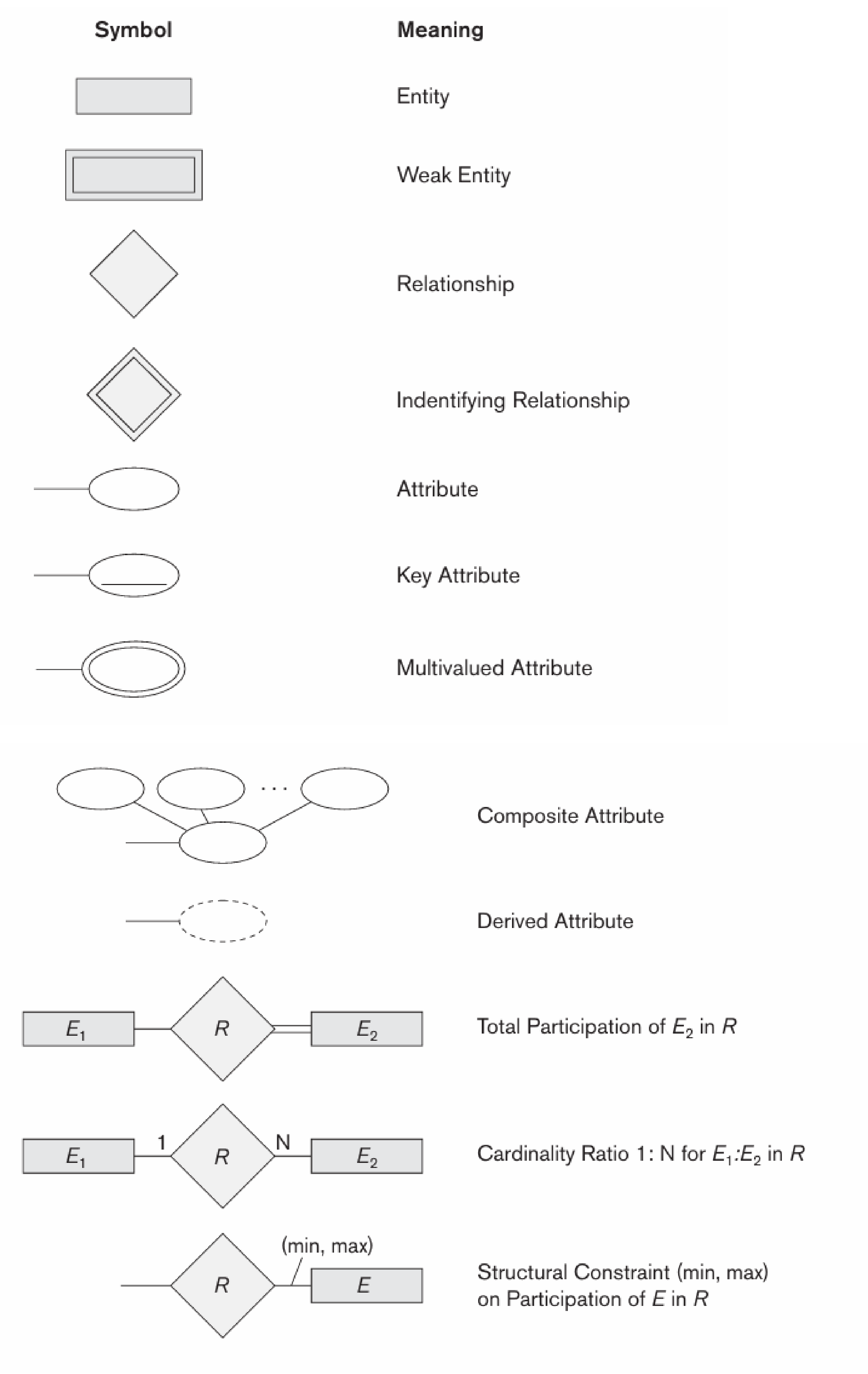
What are the advantages of a distributed database management system over a Centralized DBMS? Explain in detail.

1. **Local Data Access:** Faster access times by storing data closer to users.
2. **Load Balancing:** Workload is distributed across multiple servers.
3. **Parallel Processing:** Multiple servers handle queries simultaneously.
4. **Continuous Operation:** The system remains operational even if part of it fails.
5. **Geographical Distribution:** Data can be stored closer to users in different regions.
6. **Decentralized Control:** Each site can manage its own data independently.
7. **Tailored Solutions:** Customizable to meet local requirements and improve efficiency.
8. **Lower Cost:** Avoid large initial investments in central infrastructure.
9. **Global Access:** Facilitates data sharing and collaboration across different locations.
10. **Concurrent Access:** Multiple users can access data simultaneously.

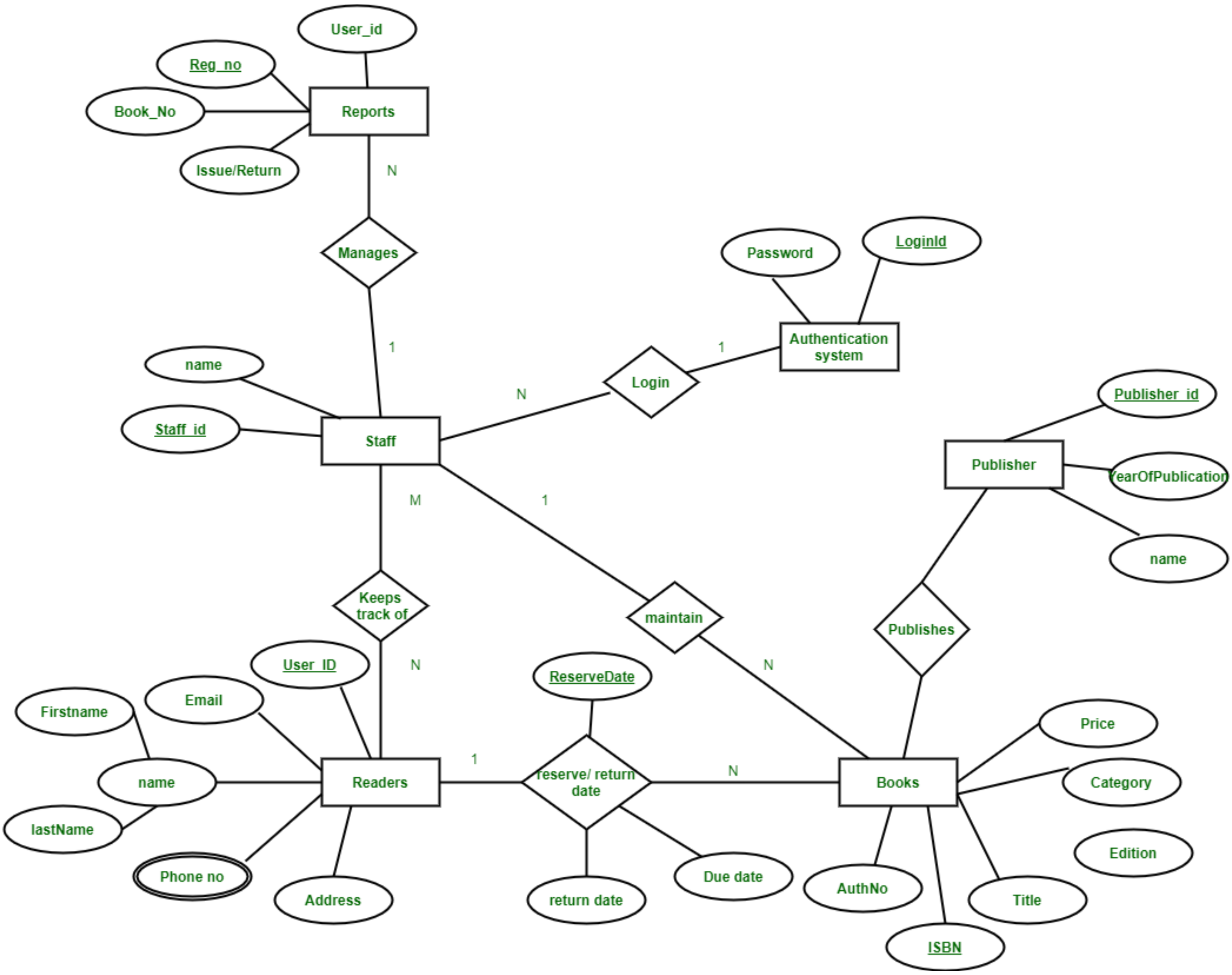
Differentiate between parallel and distributed DBMS. Quote an example for each of them.

Feature	Parallel DBMS	Distributed DBMS
Architecture	Multiple processors in a single system	Multiple independent databases across locations
Data Storage	Shared storage accessible by multiple processors	Partitioned data stored across multiple sites
Data Processing	Tasks divided and processed by different processors	Tasks distributed across various nodes, processed locally
Communication	High-speed internal bus or interconnect	Network communication with potential latency
Scalability	Vertical scaling by adding more processors	Horizontal scaling by adding more nodes
Fault Tolerance	Lower fault tolerance; central component failure impacts entire system	Higher fault tolerance; node failure doesn't affect entire system
Example	IBM Db2 with BLU Acceleration	Google Spanner

Draw an ER diagram for Library Management system and explain the various notation used.



Symbol	Meaning
	Entity
	Weak Entity
	Relationship
	Indentifying Relationship
	Attribute
	Key Attribute
	Multivalued Attribute
. . .	Composite Attribute
	Derived Attribute
E_1 — R = E_2	Total Participation of E_2 in R
E_1 1 R N E_2	Cardinality Ratio 1: N for E_1:E_2 in R
R (min, max) E	Structural Constraint (min, max) on Participation of E in R



Explain the techniques used for protecting the database against unauthorized access to a database.

1) **Authentication:**
   - Confirms user identity and grants access based on user privileges.
   - Prevents access to sensitive data using methods like biometrics (retina scans, fingerprints).
2) **Access Control:**
   - Restricts database access to authorized users by managing user accounts and login processes.
   - Tracks all user activities during login sessions.
3) **Inference Control:**
   - Prevents indirect disclosure of sensitive information by blocking inference channels.
   - Protects against identity and attribute disclosure.
4) **Flow Control:**
   - Stops unauthorized information flow to protect privacy.
   - Addresses covert channels that violate privacy policies.

5) **Database Security using Statistical Methods:**
   - Protects confidential individual values while allowing statistical queries.
   - Provides summary data without revealing personal details.
6) **Encryption:**
   - Encodes sensitive data to protect it from unauthorized access.
   - Authorized users can decode data using provided keys.

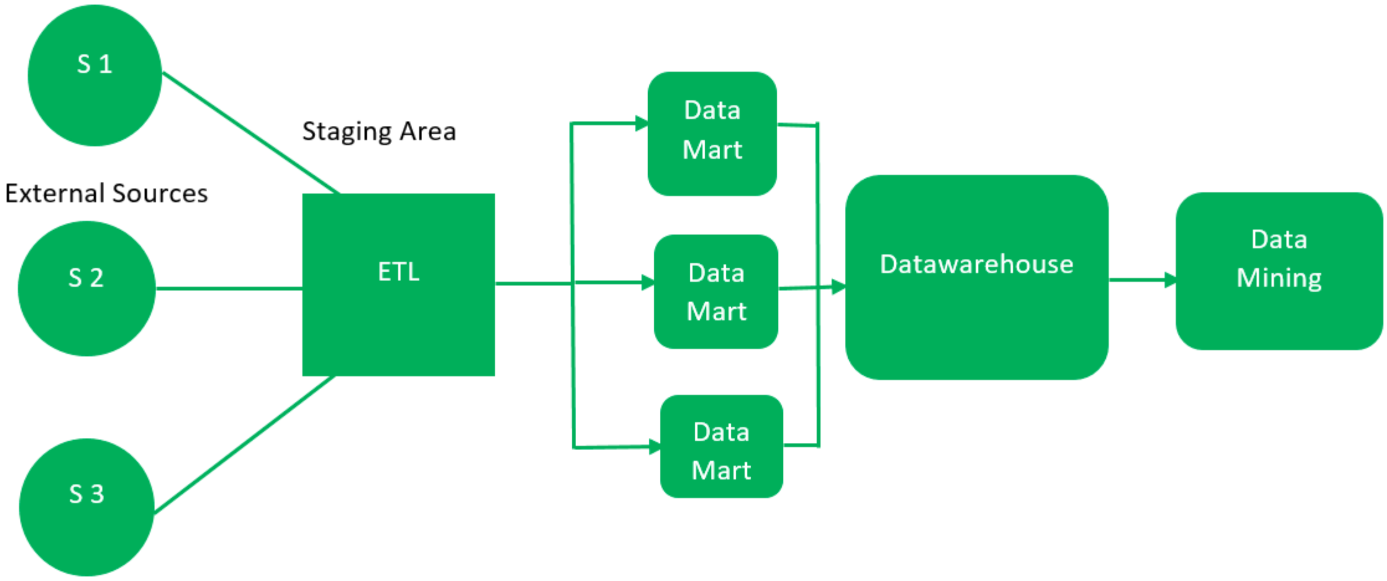
Explain the use of Data warehouse and Data mining. Discuss and explain Data Warehouse Architecture in detail.

Uses of Data Warehousing:

- **Market Basket Analysis:** Identifying products that frequently co-occur in transactions to optimize store layout and promotions.
- **Fraud Detection:** Detecting unusual patterns that may indicate fraudulent activity.
- **Customer Relationship Management (CRM):** Understanding customer behavior to improve customer service and retention.
- **Predictive Maintenance:** Predicting equipment failures before they occur to schedule proactive maintenance.

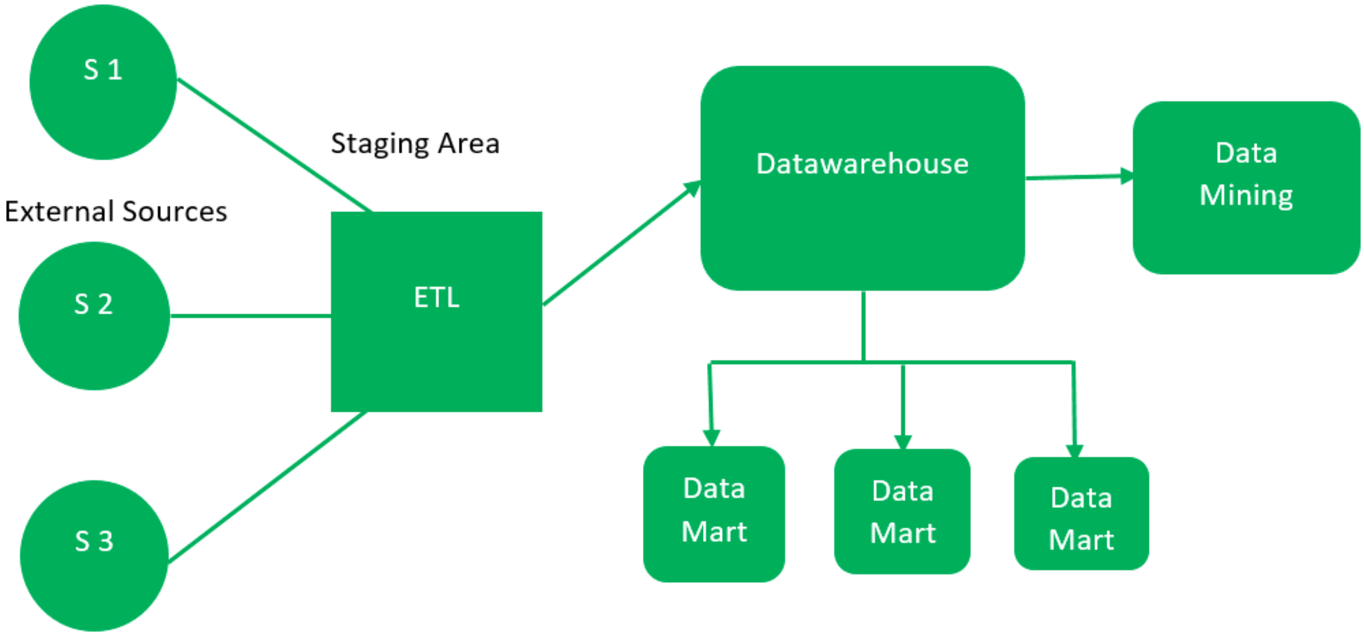
Architecture of Data Warehouse:

Bottom Up Approach:



1. First, the data is extracted from external sources.
2. Then, the data go through the staging area and loaded into data marts instead of datawarehouse. The data marts are created first and provide reporting capability. It addresses a single business area.
3. These data marts are then integrated into datawarehouse.

Top Down Approach:



Complex Data Semantic

Complex Data Semantics involves understanding and managing intricate and interconnected data relationships within a Database Management System (DBMS). This concept goes beyond simple data storage to include the meaning, context, and rules associated with the data, ensuring integrity and consistency in complex databases. Key aspects include: Hierarchical Relationships, Network Relationships, Referential Integrity, Domain Constraints, Unique Constraints, Business Rules, Composite Data Types and User-Defined Data Types. It is used in healthcare, finance and GIS.

Data Replication

Data replication is a process of making the multiple copies of database available on servers. This is done to achieve **distributed database**. This is to minimize the load on the database and provide better performance to the users.

Replicaton scheme:

1. Full replication: In full replication, the entire database is available at every site of the distributed database. This approach provides full availability and performance. In this approach, even if there is a system failure, the database availability doesn't get affected, thus this replication scheme is robust and durable.

Advantages of full replication:
1. High availability
2. Best performance
3. Full recovery in case of failure.
4. Better load balancing.

Disadvantages of full replication:
1. High storage needed

2. Data redundancy
3. Slower updates
4. Tough maintenance.

2. Partial replication: In partial replication, only the data that is frequently accessed is replicated on every site of distributed database.

Advantages of partial replication:
1. Less storage needed
2. Good performance
3. Faster updates
4. Easy Maintenance

Disadvantages of partial replication:
1. Low availability
2. Partial recovery in case of failure
3. Poor load balancing

Knowledge representation using rules

Same as association and classification rules in data mining