Sub:	Data Mining and Business Intelligence							
Date:	03/08/23	Duration:	90 min's	Max Marks:	50	Sem:	II

Scheme

1 Question-1
Explain in detail the building blocks of Data Warehouse.
Each block - 2 marks
Diagram - 2marks
2 Question-2
What is Data warehouse? Explain the various Data warehouse models.
Each model - 3 marks
Virtual warehouse - 1 mark
Each model-3 marks
3 PART II
Question-3
Explain the Top-Down, Bottom-Up and Combined approach in Data Warehouse.
Each layer - 3 marks
Diagram - 1 marks

4
Question-4
Describe the 3tier data warehouse architecture.
Diagram - 4 marks
Each tier - 2 marks
5 PART III
Question-5
Explain the OLAP Operations with examples.
Each 2 marks

6 Question-6
What are the characteristics of a Data warehouse? What is metadata in a Data warehouse?
Characteristics - 4marks
Metadata - 6 marks

7 PART IV
Question-7
What are the various types of Data Marts?
Explain.
Definition 1 marks
Each 3 marks
8 Question-8
Compare OLTP and OLAP.
Each point 2 marks

9 PART V
Question-9
What are the 3 types of Schemas in a multidimensional data model? Explain.

10 Question-10
Explain with a neat diagram the Knowledge discovery of data.

SOLUTION

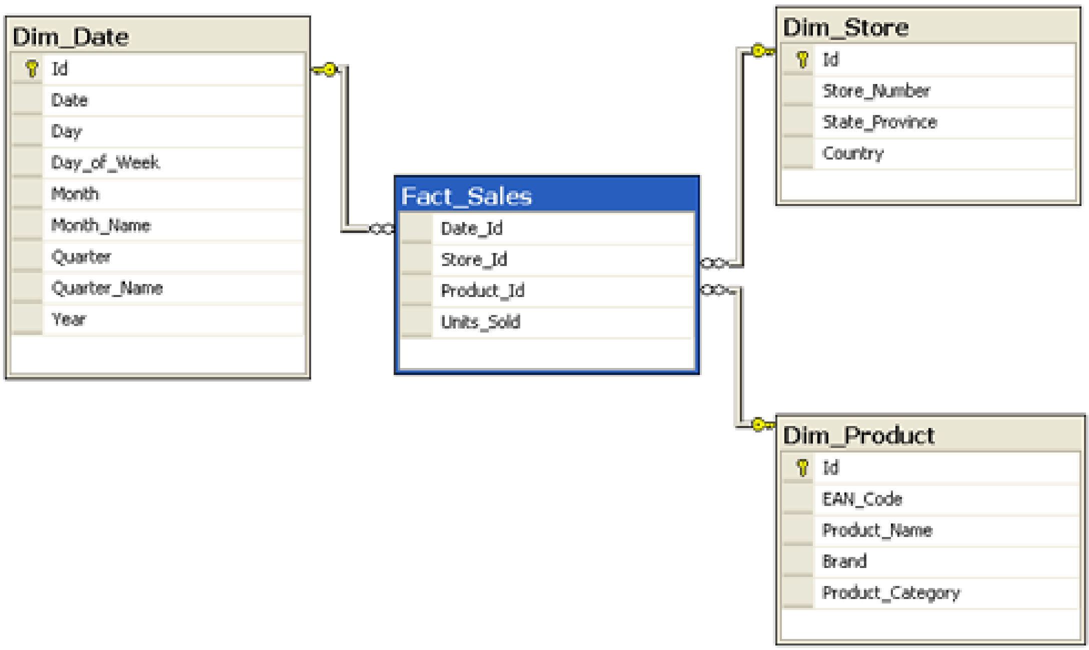
Note : Answer FIVE FULL Questions, choosing ONE full question from each Module

PART I

1 **Question-1**

Explain in detail the building blocks of Data Warehouse.

- A) A data warehouse is a relational database that is designed for query and analysis.
- It separates an analysis workload from a transaction workload and enables an organization to consolidate data from several sources.
- B) Dimensional modeling: It is developed to be oriented around query performance and ease of use. The dimensional modeling handle approach is at a logical level.
  - ✔ Facts or Business measurement-numeric values
  - ✔ Dimensions or Descriptors specify the facts- text values
- C)Star Scheme: The fact table is at the center of the schema surrounded by dimensions tables.
  - ✔ Eg. At the center of the schema there is fact table FACT-SALES.
  - ✔ The fact table is sourrounded by the dimension tables Dim-Data, Dim-Store, Dim-Product.
- D)Fact Table: It is a dimensional model in data warehouse design. Facts are also known as measurements.
  - ✔ Types of Fact table are Transactional, periodic and accumulating tables.



- Transactional –Transactional fact table is the most basic one that each grain associated with it indicated as “one row per line in a transaction”, e.g.,Price- every line item appears on an invoice.
- Periodic snapshots – Periodic snapshots fact table stores the data that is a snapshot in a period of time. Ex. Sales period
- Accumulating snapshots – The accumulating snapshots fact table describes the activity of a business process that has a clear beginning and end. Eg. Purchasing: Requisition, Purchase order, Vendor Invoice, Delivery, Payment.

OR

2 **Question-2**

What is Data warehouse? Explain the various Data warehouse models.

- **Enterprise Data Warehouse**:
  - Enterprise Data Warehouse is a **centralized warehouse**, which provides **decision support service across the enterprise**.
  - It offers a **unified approach to organizing and representing dat**a.
  - It also provides the **ability to classify data according to the subject** and give access according to those divisions.
- **Operational Data Store**:
  - Operational Data Store, also called ODS, is data store required when neither data warehouse nor OLTP systems support organizations reporting needs.
  - It is widely preferred for **routine activities like storing records.**.
  - In ODS, Data warehouse is refreshed in real time.
- **Data Mart**:
  - A Data Mart is a subset of the data warehouse.
  - It specially designed for specific segments like sales, finance, sales, or finance.
  - In an independent data mart, data can collect directly from sources.

- **Virtual warehouse:**
  - A virtual warehouse is a set of views over operational databases.

  For efficient query processing, only some of the possible summary views may be materialized.

PART II

3 **Question-3**

Explain the Top-Down, Bottom-Up and Combined approach in Data Warehouse.

- **Top Down Approach**
  - The top-down approach starts with the overall design and planning.
  - It is useful in cases where the technology is mature and well known, and where the business problems that must be solved are clear and well understood.

- **Bottom up Approach**
  - The bottom-up approach starts with experiments and prototypes.
  - This is useful in the early stage of business modeling and technology development.
  - It allows an organization to move forward at considerably less expense and to evaluate the benefits of the technology before making significant commitments.

- **Combined Approach**
  - In the combined approach, an organization can exploit the planned and strategic nature of the top-down approach while retaining the rapid implementation and opportunistic application of the bottom-up approach.

The warehouse design process consists of the following steps:

- Choose a business process to model, for example, orders, invoices, shipments, inventory, account administration, sales, or the general ledger.

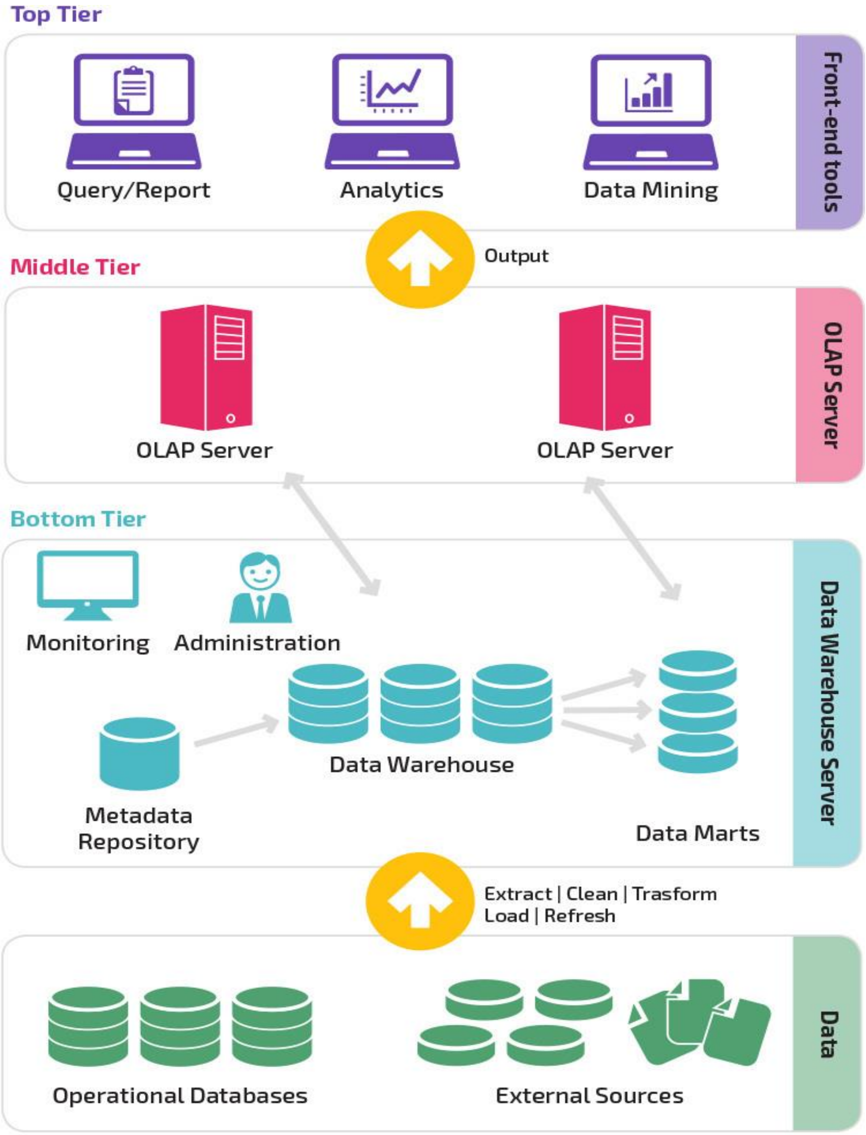
- If the business process is organizational and involves multiple complex object collections, a data warehouse model should be followed. However, if the process is departmental and focuses on the analysis of one kind of business process, a data mart model should be chosen.

- Choose the grain of the business process. The grain is the fundamental, atomic level of data to be represented in the fact table for this process, for example, individual transactions, individual daily snapshots, and so on.
- Choose the dimensions that will apply to each fact table record. Typical dimensions are time, item, customer, supplier, warehouse, transaction type, and status.
- Choose the measures that will populate each fact table record. Typical measures are numeric additive quantities like dollars sold and units sold.

OR

4 **Question-4**

Describe the 3tier data warehouse architecture.



- Bottom tier:
  - The bottom tier is a warehouse data warehouse server that is almost always a relational database system.
  - Back-end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources.
  - These tools and utilities perform data extraction, cleaning, and transformation, as

well as load and refresh functions to update the data warehouse.

- The data are extracted using application program interfaces known as gateways.
- Examples of gateways include ODBC (Open Database Connection) and OLEDB (Open Linking and Embedding for Databases) by Microsoft and JDBC (Java Database Connection).
- This tier also contains a metadata repository, which stores information about the data warehouse and its contents.

▪ Middle tier:

- OLAP servers use the Snowflakes schema.
- The middle tier is an OLAP (Online Analytical Processing Server) that is typically implemented using either
    - A relational OLAP (ROLAP) model, that is, an extended relational DBMS that maps operations on multidimensional data to standard relational operations or,
    - A multidimensional OLAP (MOLAP) model, that is, a special-purpose server that directly implements multidimensional data and operations.

▪ Top tier:

The top tier is a front-end client layer, which contains query and reporting tools, analysis tools, and/or data mining tools.

5

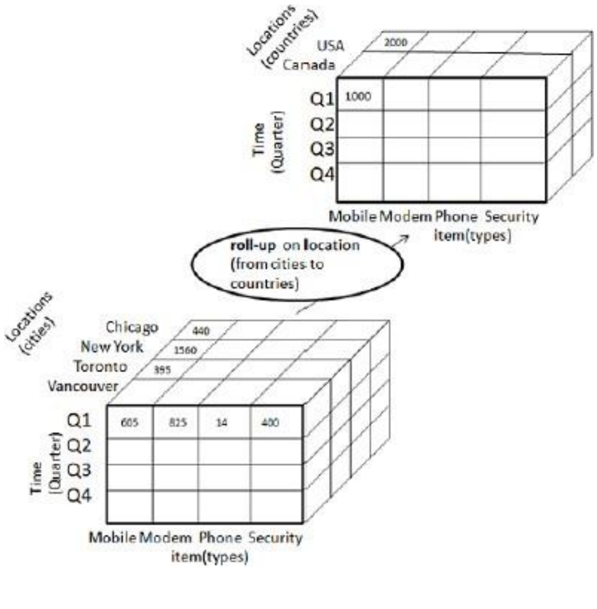
PART III

Question-5

Explain the OLAP Operations with examples.

1. **Roll-up**

- Roll-up performs aggregation on a data cube in any of the following ways:
    - By climbing up a concept hierarchy for a dimension
    - By dimension reduction
- The following diagram illustrates how roll-up works.



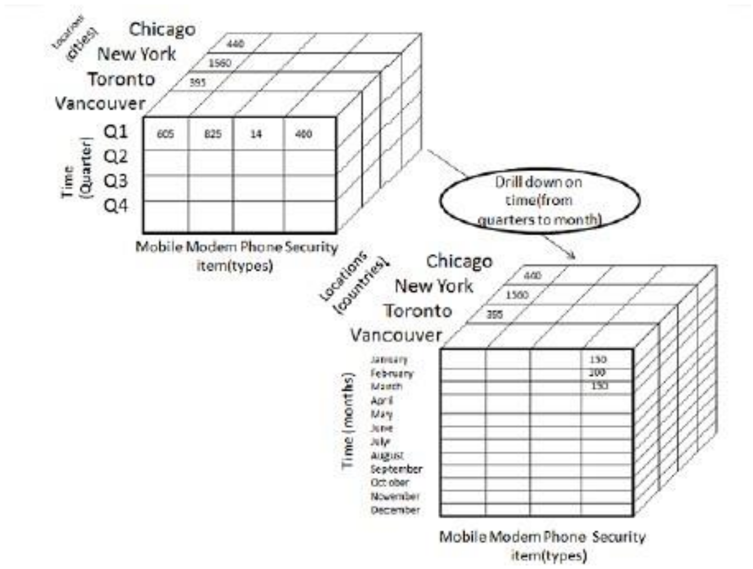
- Roll-up is performed by climbing up a concept hierarchy for the dimension location.
- Initially the concept hierarchy was "street < city < province < country".
- On rolling up, the data is aggregated by ascending the location hierarchy from the level

of city to the level of country.

- The data is grouped into cities rather than countries.
- When roll-up is performed, one or more dimensions from the data cube are removed.

2. **Drill-down**

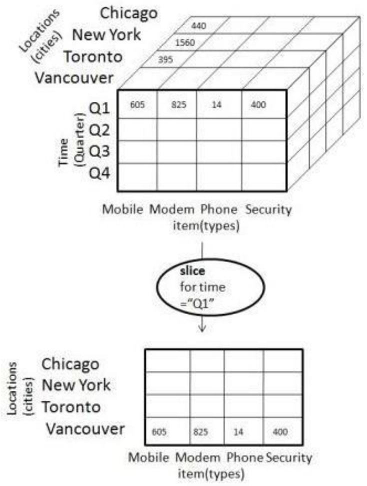
- Drill-down is the reverse operation of roll-up. It is performed by either of the following ways:
    - By stepping down a concept hierarchy for a dimension
    - By introducing a new dimension.
- The following diagram illustrates how drill-down works:



- Drill-down is performed by stepping down a concept hierarchy for the dimension time.
- Initially the concept hierarchy was "day < month < quarter < year."
- On drilling down, the time dimension is descended from the level of quarter to the level of month.
- When drill-down is performed, one or more dimensions from the data cube are added.
- It navigates the data from less detailed data to highly detailed data.

3. **Slice**

- The slice operation selects one particular dimension from a given cube and provides a new subcube.



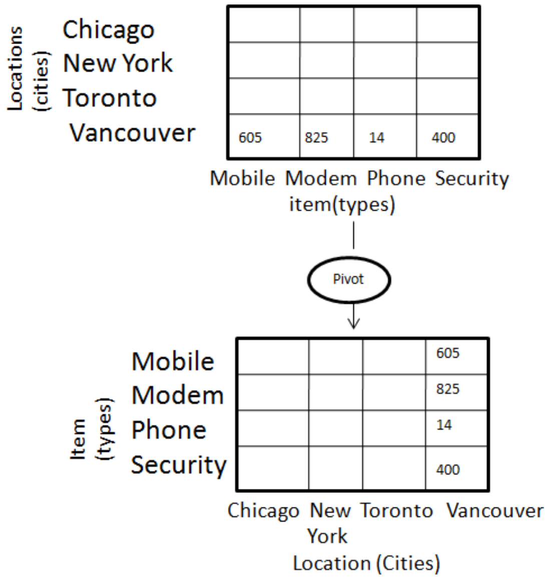
- Consider the following diagram that shows how slice works.

4. **Dice**

- Dice selects two or more dimensions from a given cube and provides a new sub-cube.
- Consider the following diagram that shows the dice operation.
- The dice operation on the cube based on the following selection criteria involves three dimensions.
  - (location = "Toronto" or "Vancouver")
  - (time = "Q1" or "Q2")
  - (item =" Mobile" or "Modem")

5. **Pivot**

- The pivot operation is also known as rotation.
- It rotates the data axes in view in order to provide an alternative presentation of data.
- Consider the following diagram that shows the pivot operation.
- In this the item and location axes in 2-D slice are rotated.



6 **Question-6**

What are the characteristics of a Data warehouse? What is metadata in a Data warehouse?

- **Subject-oriented:**

- A data warehouse is organized around major subjects, such as customer, supplier, product, and sales.
- Rather than concentrating on the day-to-day operations and transaction processing of an organization, a data warehouse focuses on the modeling and analysis of data for decision makers.
- Data warehouses typically provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

- **Integrated:**
  - A data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and on-line transaction records.
  - Data cleaning and data integration techniques are applied to ensure consistency in naming conventions, encoding structures, attribute measures, and so on.
- **Time-variant:**
  - Data are stored to provide information from a historical perspective (e.g., the past 5–10 years).
  - Every key structure in the data warehouse contains, either implicitly or explicitly, an element of time.
- **Nonvolatile:**
  - A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment.
  - Due to this separation, a data warehouse does not require transaction processing, recovery, and concurrency control mechanisms.
  - It usually requires only two operations in data accessing: initial loading of data and access of data.

- Metadata are data about data. When used in a data warehouse, metadata are the data that define warehouse objects.
- Metadata are created for the data names and definitions of the given warehouse.
- Additional metadata are created and captured for time stamping any extracted data, the source of the extracted data, and missing fields that have been added by data cleaning or integration processes.

A metadata repository should contain the following:

- A description of the structure of the data warehouse, which includes the warehouse schema, view, dimensions, hierarchies, and derived data definitions, as well as data mart locations and contents.

- Operational metadata, which include data lineage (history of migrated data and the sequence of transformations applied to it), currency of data (active, archived, or purged), and monitoring information (warehouse usage statistics, error reports, and audit trails).

- The algorithms used for summarization, which include measure and dimension definition algorithms, data on granularity, partitions, subject areas, aggregation, summarization and predefined queries and reports.

- The mapping from the operational environment to the data warehouse, which includes source databases and their contents, gateway descriptions, data partitions, data extraction, cleaning, transformation rules and defaults, data refresh and purging rules, and security (user authorization and access control).

- Data related to system performance, which include indices and profiles that improve data access and retrieval performance, in addition to rules for the timing and scheduling of refresh, update, and replication cycles.

- Business metadata, which include business terms and definitions, data ownership information, and charging policies.

7

PART IV

Question-7

What are the various types of Data Marts? Explain.

- Data marts contain a subset of organization-wide data that is valuable to specific groups of people in an organization.
- A data mart contains only those data that is specific to a particular group.
- Data marts improve end-user response time by allowing users to have access to the specific type of data they need to view most often by providing the data in a way that supports the collective view of a group of users.

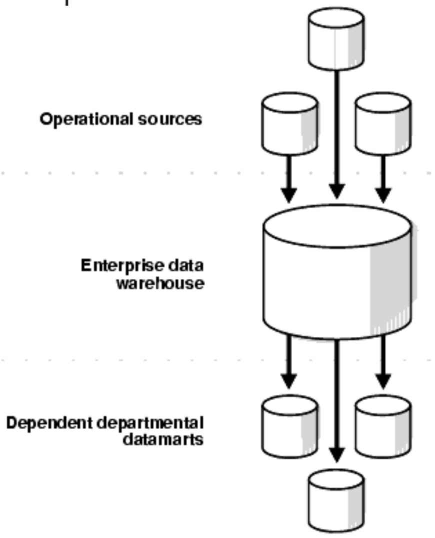
- A data mart is basically a condensed and more focused version of a data warehouse that reflects the regulations and process specifications of each business unit within an organization.

- Each data mart is dedicated to a specific business function or region.

- For example, the marketing data mart may contain only data related to items, customers, and sales. Data marts are confined to subjects.

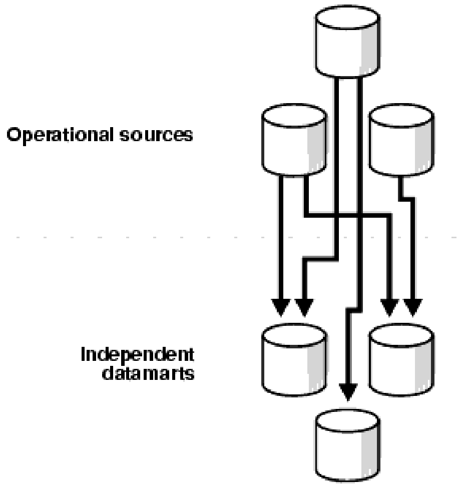
1. Dependent Data Marts

- A dependent data mart allows you to unite your organization's data in one data warehouse.
- This gives you the usual advantages of centralization.
- Figure illustrates a dependent data mart.



2. Independent Data Marts

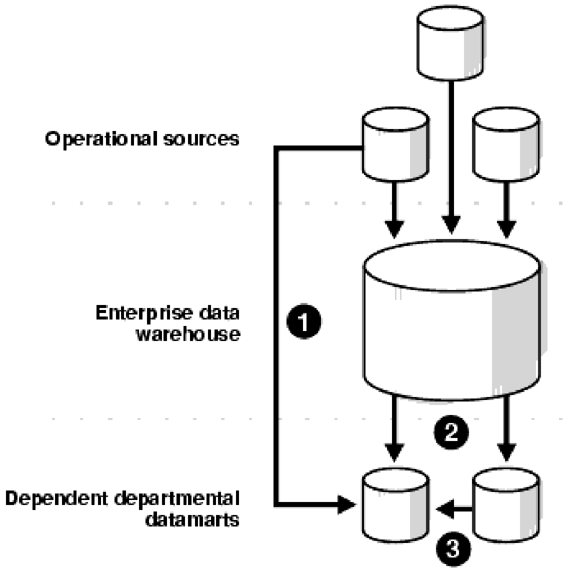
- An independent data mart is created without the use of a central data warehouse.
- This could be desirable for smaller groups within an organization.



- Figure illustrates an independent data mart.

3. Hybrid Data Marts

- A hybrid data mart allows you to combine input from sources other than a data warehouse.
- This could be useful for many situations, especially when you need ad hoc integration, such as after a new group or product is added to the organization.



- Figure illustrates a hybrid data mart.

Question-8
Compare OLTP and OLAP.

- OLAP is characterized by relatively **low volume of transactions**.
- Queries are often **very complex and involve aggregations**.
- For OLAP systems a **response time is an effectiveness measure**.
- OLAP applications are widely used by Data Mining techniques.
- In OLAP database there is **aggregated, historical data, stored in multi-dimensional** schemas (usually star schema).

- OLTP is characterized by a **large number of short on-line transactions** (INSERT, UPDATE, DELETE).
- The main emphasis for OLTP systems is put on very fast query processing, maintaining data integrity in multi-access environments and an effectiveness measured by number of transactions per second.
- In OLTP database, **there is detailed and current data**, and schema used to store transactional databases is the entity model (usually 3NF).

8

Functionality	**OLTP**	**OLAP**
Characteristic	Operational processing informational processing	Transaction Analysis
Orientation	Transaction	Analysis
User	Clerk, DBA, database professional	Knowledge worker (e.g., manager, executive, analyst)
Function	day-to-day operations	long-term informational requirements, decision support
DB design	ER based, **application-oriented**	Star/snowflake, **subject-oriented**
Data	Current; guaranteed up-to-date	Historical; accuracy maintained over time
Summarization	Primitive, highly detailed	Summarized, consolidated
View	Detailed, flat relational	Summarized, multidimensional

Unit of work	Short, simple transaction	Complex query
Access	Read/write	Mostly read

PART V

9

Question-9

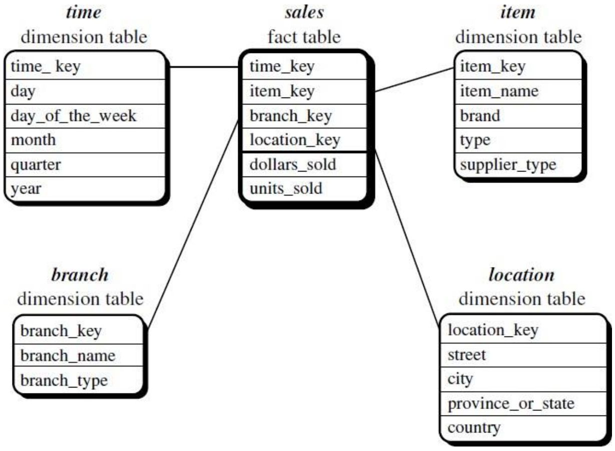
What are the 3 types of Schemas in a multidimensional data model? Explain.

Star schema: The most common modeling paradigm is the star schema, in which the data warehouse contains,

(1) a large central table (fact table) containing the bulk of the data, with no redundancy, and

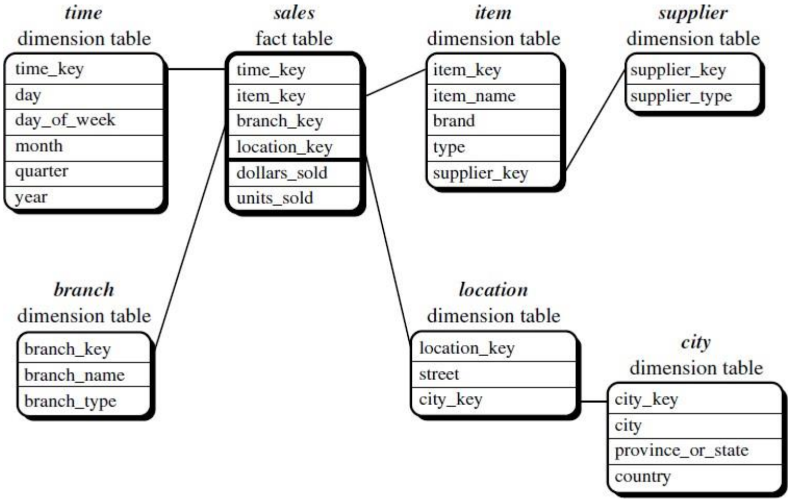
(2) a set of smaller attendant tables (dimension tables), one for each dimension.

- The schema graph resembles a starburst, with the dimension tables displayed in a radial pattern around the central fact table.



```
define cube sales star [time, item, branch, location]:
dollars sold = sum(sales in dollars), units sold = count(*)
define dimension time as (time key, day, day of week, month,
quarter, year) define dimension item as (item key, item name,
brand, type, supplier type) define dimension branch as (branch
key, branch name, branch type)
define dimension location as (location key, street, city, province or state, country)
```

Snowflake shema: The major **difference between the snowflake and star schema** models is that the dimension tables of the snowflake model may be kept in normalized form to reduce redundancies. Such a table is easy to maintain and saves storage space.

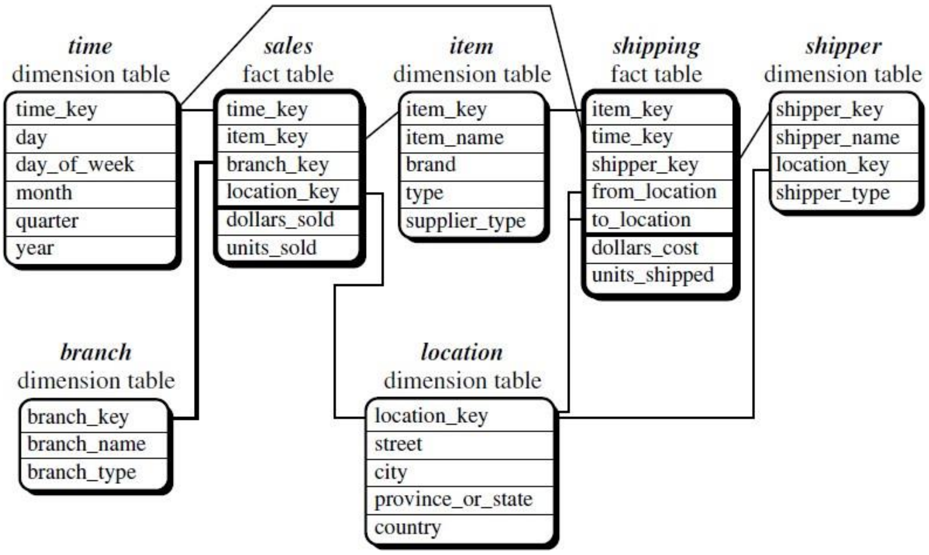


- However, this saving of space is negligible in comparison to the typical magnitude of the fact table. Furthermore, the snowflake structure can reduce the effectiveness of browsing, since more joins will be needed to execute a query.

- Hence, although the snowflake schema reduces redundancy, it is not as popular as the star schema in data warehouse design.
  define cube sales snowflake [time, item,
  branch, location]: dollars sold = sum(sales in
  dollars), units sold = count(*)
  define dimension time as (time key, day, day of week, month,
  quarter, year) define dimension item as (item key, item name,
  brand, type, supplier (supplier key, supplier type))
  define dimension branch as (branch key, branch name,
  branch type) define dimension location as (location
  key, street, city
  (city key, city, province or state, country))

Fact constellation: Sophisticated applications may require multiple fact tables to *share* dimension tables.

- This kind of schema can be viewed as a collection of stars, and hence is called a galaxy schema or a fact constellation.
- A fact constellation schema allows dimension tables to be shared between fact tables.
- For example, the dimensions tables for *time, item*, and *location* are shared between both the *sales* and
  shipping fact tables.



define cube sales [time, item, branch, location]:
dollars sold = sum(sales in dollars), units sold = count(*)
define dimension time as (time key, day, day of week, month,
quarter, year) define dimension item as (item key, item name,
brand, type, supplier type) define dimension branch as (branch
key, branch name, branch type)
define dimension location as (location key, street, city, province or
state, country)
define cube shipping [time, item, shipper, from location,
to location]: dollars cost = sum(cost in dollars), units

shipped = count(*)
define dimension time as time
in cube sales define dimension
item as item in cube sales
define dimension shipper as (shipper key, shipper
name, location as location in cube sales, shipper
type)
define dimension from location as location
in cube sales define dimension to location
as location in cube sales

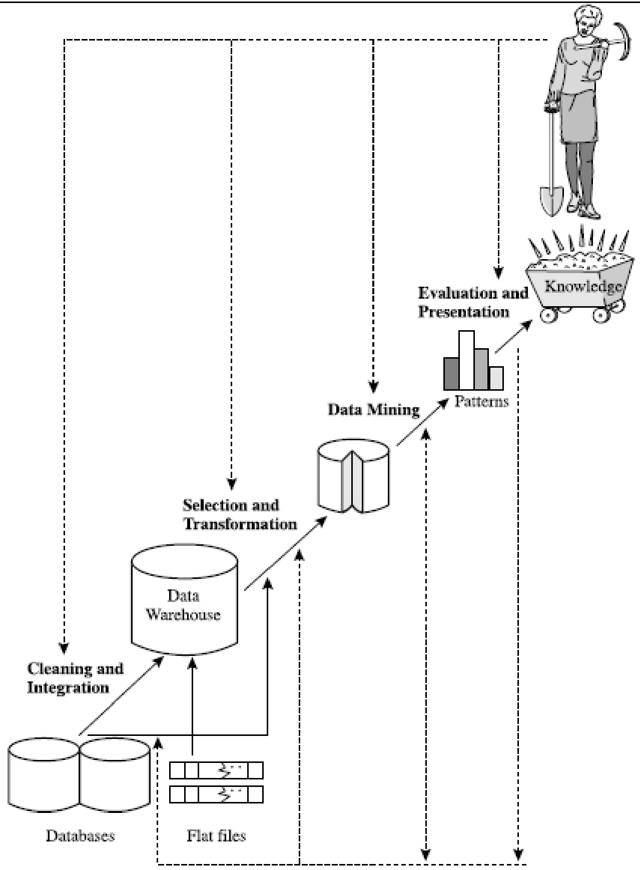
10

Question-10

Explain with a neat diagram the Knowledge discovery of data.

KDD (Knowledge Discovery from Data) Process

- KDD stands for knowledge discoveries from database. There are some pre-processing operations which are required to make pure data in data warehouse before use that data for Data Mining processes.
- A view data mining as simply an essential step in the process of knowledge discovery. Knowledge discovery as a process is depicted in Figure 2 and consists of an iterative sequence of the following steps:
  - ✓ **Data cleaning:** To remove noise and inconsistent data.
  - ✓ **Data integration:** where multiple data sources may be combined.
  - ✓ **Data selection:** where data relevant to the analysis task are retrieved from the database.
  - ✓ **Data transformation**: where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance.
  - ✓ **Data mining**: An essential process where intelligent methods are applied in order to extract data patterns.
  - ✓ **Pattern evaluation**: To identify the truly interesting patterns representing knowledge based on some interestingness measures.
  - ✓ **Knowledge presentation**: where visualization and knowledge representation techniques are used to present the mined knowledge to the user.



- KDD refers to the overall process of discovering useful knowledge from data. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data prior to the data mining step.
- Data mining refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process.
- Objective of Pre-processing on data is to remove noise from data or to remove redundant data.
- There are mainly 4 types of Pre-processing Activities included in KDD Process that is shown in fig. as Data cleaning, Data integration, Data transformation, Data reduction.