CMR
INSTITUTE OF
TECHNOLOGY

USN | | | | | | | | | | |

Internal Assessment Test 1 – July 2022

Data Mining with Business Intelligence							Sub Code:	20MCA252
29/07/22	Duration:	90 min's	Max Marks:	50	Sem:	II	Branch:	MCA

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

PART I

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.

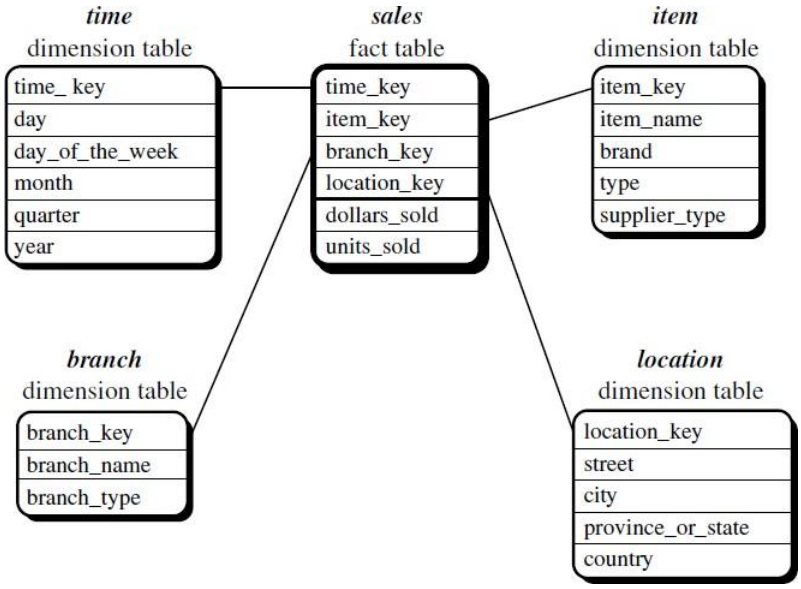
B) Dimensional modeling:
It is developed to be oriented around query performance and ease of use. The dimensional modeling handle approach is at a logical level.

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

Question-2

What is Data Warehouse? Explain it with Key Feature.

- Data warehousing provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions.
- A data warehouse refers to a database that is maintained separately from an organization's operational databases.
- Data warehouse systems allow for the integration of a variety of application systems.
- They support information processing by providing a solid platform of consolidated historical data for analysis.
- According to William H. Inmon, a leading architect in the construction of data warehouse systems, "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision making process"
- The four keywords, subject-oriented, integrated, time-variant, and nonvolatile, distinguish data warehouses from other data repository systems, such as relational database systems, transaction processing systems, and file systems.

  - **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.

PART II
Question-3
Explain the Top-Down, Bottom-Up and Combined approach in Data Warehouse.

A data warehouse can be built using a *top-down approach*, a *bottom-up approach*, or a *combination of both*.

- **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

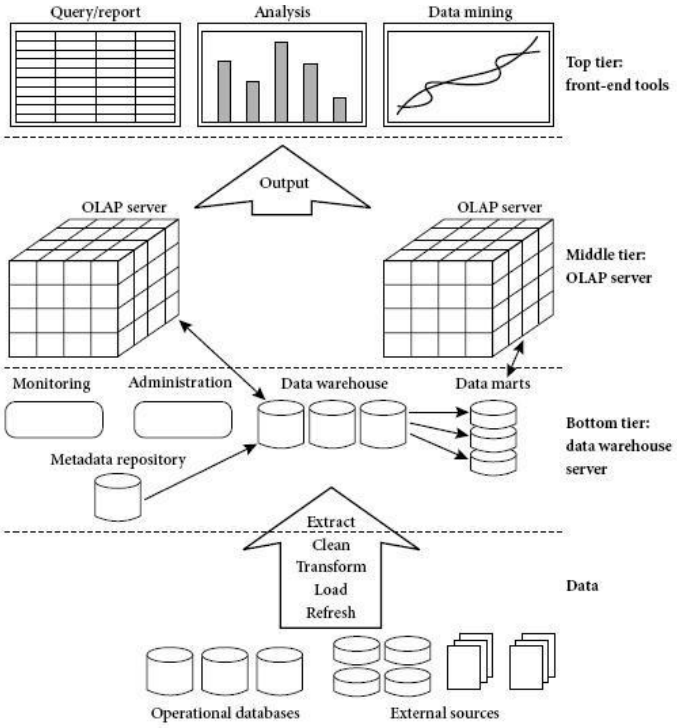
represented in the fact table for this process, for example, individual transactions, individual daily snapshots, and so on.

- 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

Question-4

Explain with neat diagram the 3 tier architecture of a data warehouse .



Bottom tier:

- The **bottom tier** is a warehouse **database 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.
- A gateway is supported by the underlying DBMS and allows client programs to generate SQL code to be executed at a server.
- 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:

- The middle tier is an OLAP 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.

From the architecture point of view, there are **three data warehouse models:**

1. Enterprise warehouse:

- An enterprise warehouse collects all of the information about subjects spanning the entire organization.
- It provides corporate-wide data integration, usually from one or more operational systems or external information providers, and is cross-functional in scope.
- It typically contains detailed data as well as summarized data,
- It can range in size from a few gigabytes to hundreds of gigabytes, terabytes, or beyond.

2. Data mart:

- A data mart contains a subset of corporate-wide data that is of value to a specific group of users.

3. Virtual warehouse:

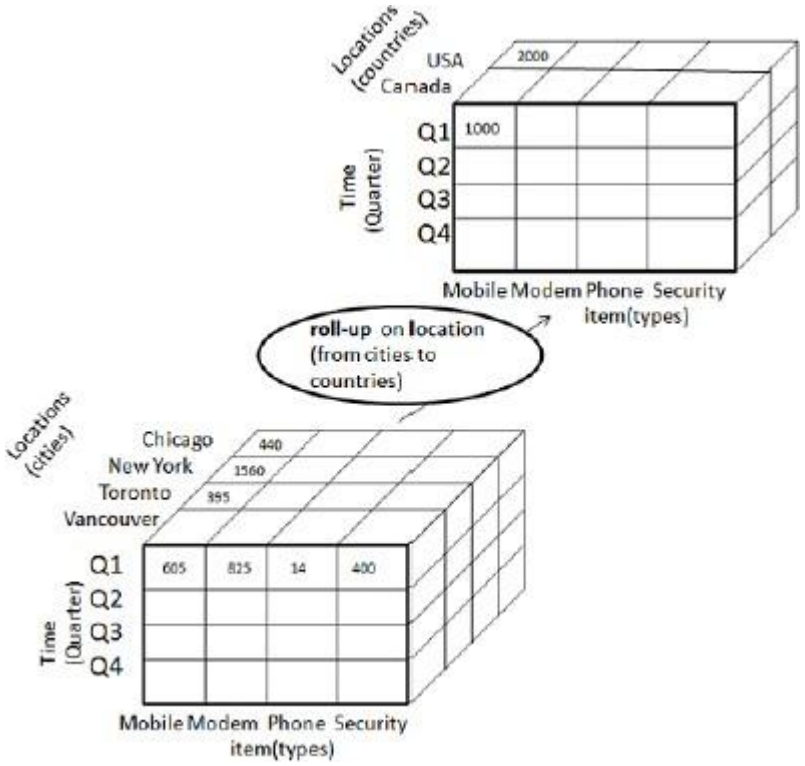
- A virtual warehouse is a set of views over operational databases.

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

PART III
Question-5
Explain the OLAP Operations with examples.

1. **Roll-up**

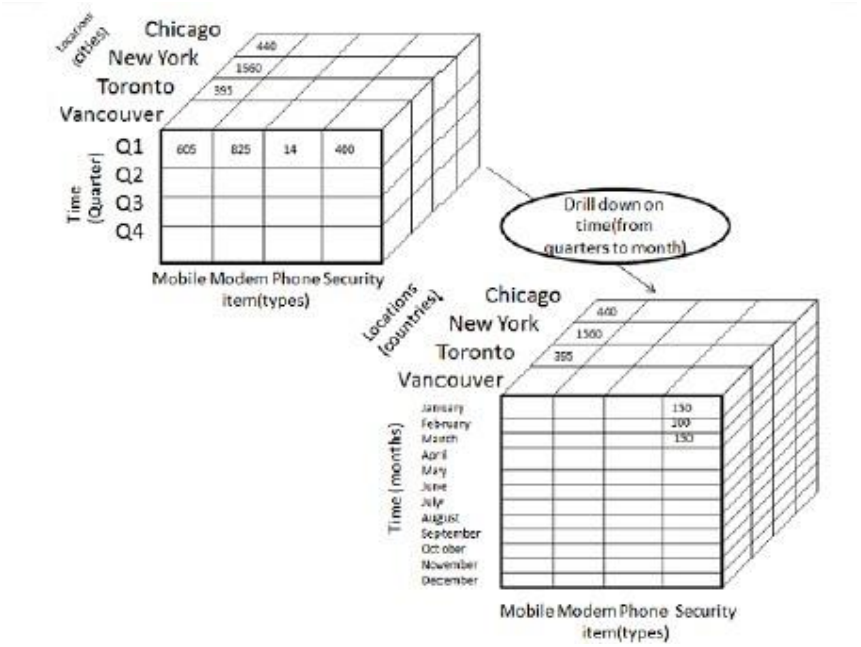
- Roll-up performs aggregation on a data cube in any of the following ways:
  - By climbing up a concept hierarchy for a dimension
  - By dimension reduction
- The following diagram illustrates how roll-up works.



- 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

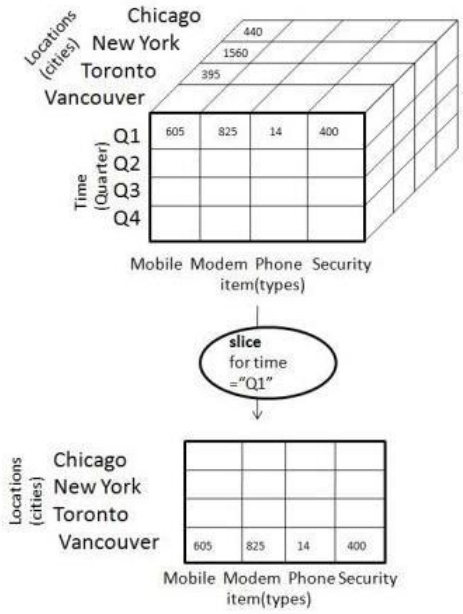
- Drill-down is the reverse operation of roll-up. It is performed by either of the following ways:
  - By stepping down a concept hierarchy for a dimension
  - By introducing a new dimension.
- The following diagram illustrates how drill-down works:



- 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

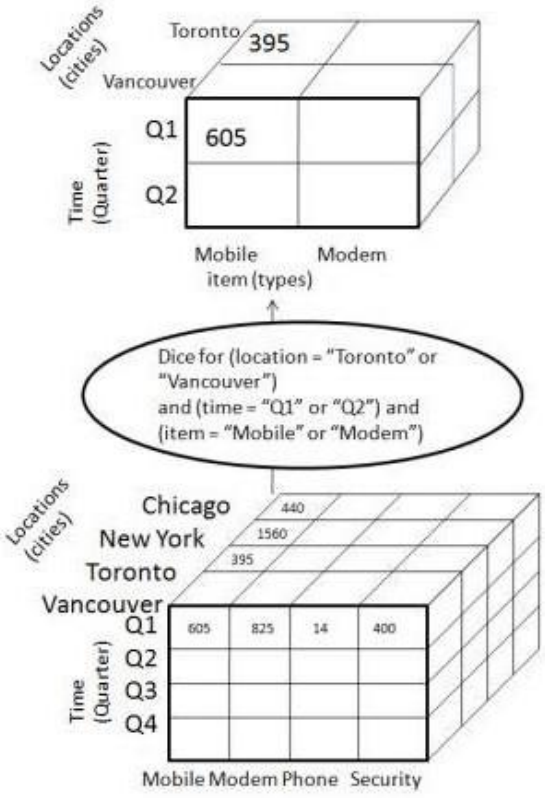
- The slice operation selects one particular dimension from a given cube and provides a new subcube.
- Consider the following diagram that shows how slice works.



- Here Slice is performed for the dimension "time" using the criterion time = "Q1".
- It will form a new sub-cube by selecting one or more dimensions.

4. Dice

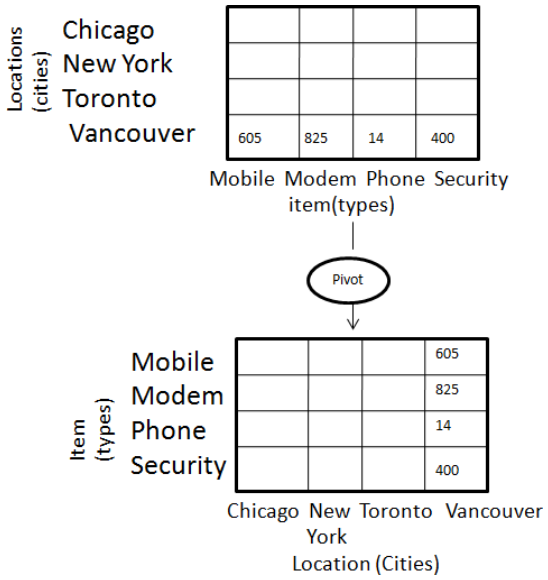
- Dice selects two or more dimensions from a given cube and provides a new sub-cube.
- Consider the following diagram that shows the dice operation.



- 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.



OR

Question-6

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

- 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.

PART IV

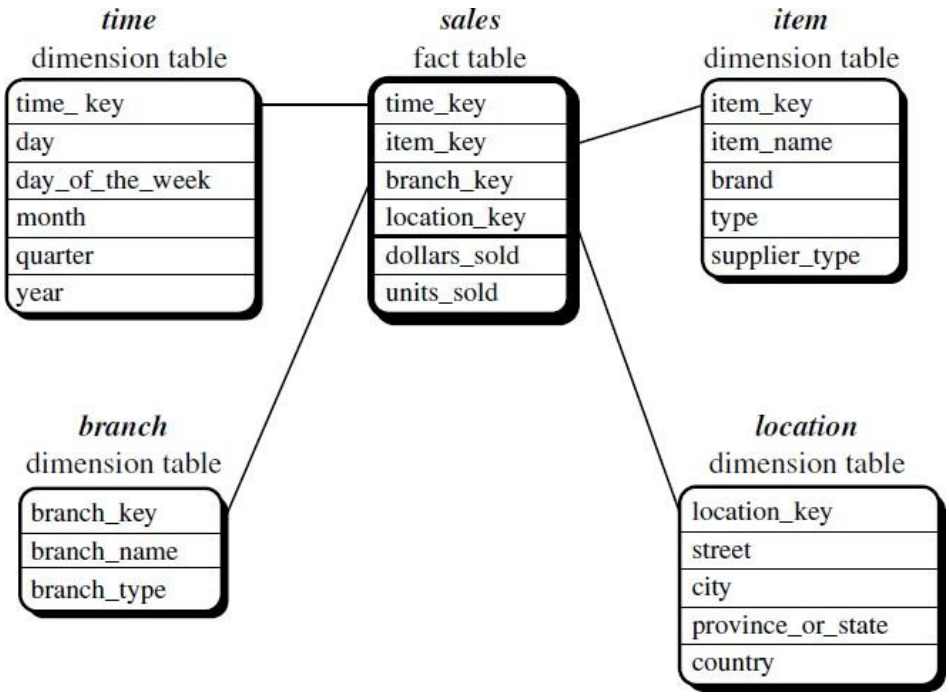
Question-7

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.
- DMQL code for star schema can be written as follows:



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.

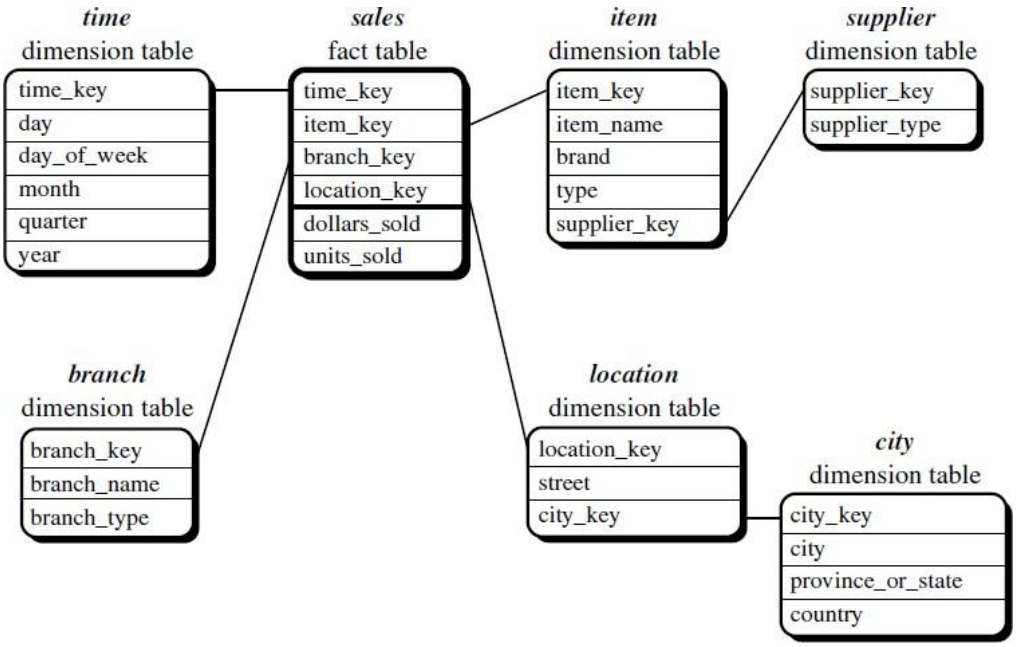
- However, this saving of space is negligible in comparison to the typical magnitude of the fact table. Furthermore, the snowflake structure can reduce the effectiveness of browsing, since more joins will be needed to execute a query.
- Hence, although the snowflake schema reduces redundancy, it is not as popular as the star schema in data warehouse design.
- DMQL code for star schema can be written as follows:

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))

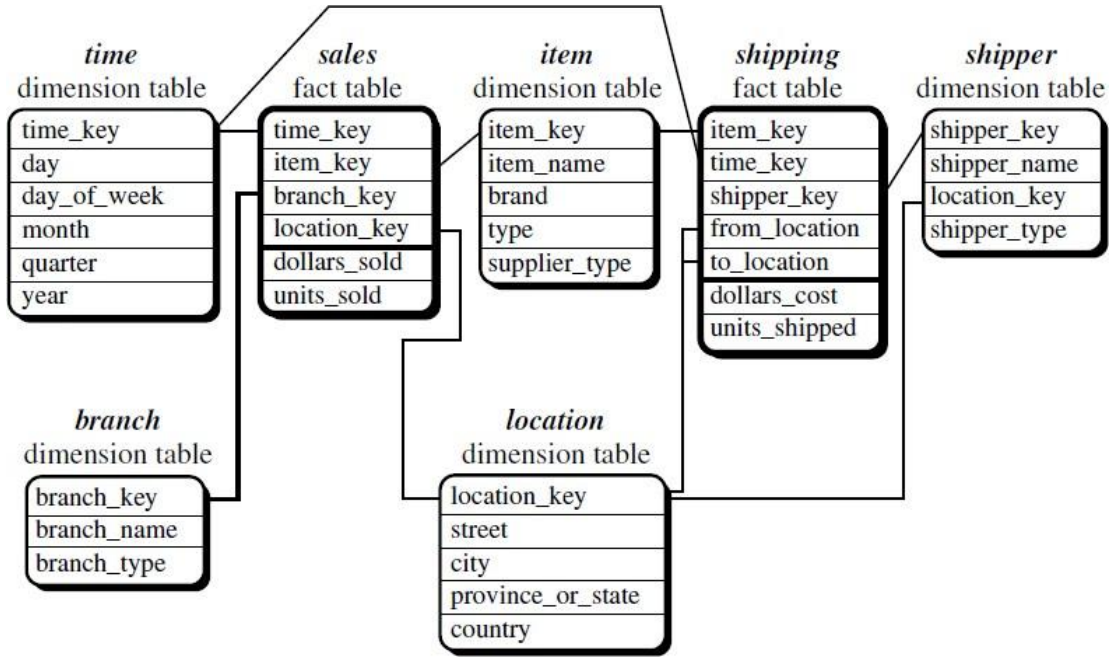
define dimension branch as (branch key, branch name, branch type)
define dimension location as (location key, street, city

(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.



OR

Question-8

Explain the terms Dimensional modeling, virtual warehouse, data cube, ROLAP, and MOLAP.

1. **Relational OLAP**

- ROLAP servers are placed between relational back-end server and client front-end tools.
- To store and manage warehouse data, ROLAP uses relational or extended-relational DBMS.
- ROLAP includes the following:
  - Implementation of aggregation navigation logic.
  - Optimization for each DBMS back end.
  - Additional tools and services.

2. **Multidimensional OLAP**

- MOLAP uses array-based multidimensional storage engines for multidimensional views of data.
- With multidimensional data stores, the storage utilization may be low if the data set is sparse.
- Many MOLAP server use two levels of data storage representation to handle dense and sparse data sets.

3. **Hybrid OLAP (HOLAP)**

- Hybrid OLAP is a combination of both ROLAP and MOLAP.
- It offers higher scalability of ROLAP and faster computation of MOLAP.
- HOLAP servers allows to store the large data volumes of detailed information.
- The aggregations are stored separately in MOLAP store.

4. **Specialized SQL Servers**

- Specialized SQL servers provide advanced query language and query processing support for SQL queries over star and snowflake schemas in a read-only environment.

PART V

Question-9

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.

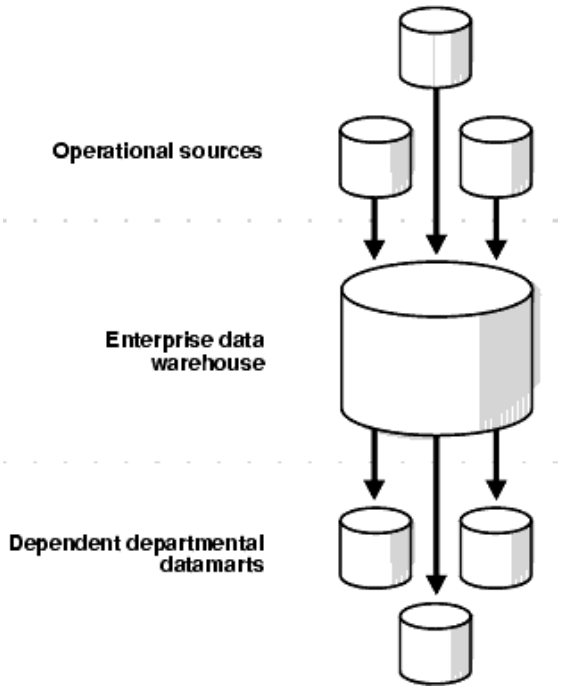
- Each data mart is dedicated to a specific business function or region.

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

- Three basic types of data marts are dependent, independent, and hybrid.
- The categorization is based primarily on the data source that feeds the data mart.
- Dependent data marts draw data from a central data warehouse that has already been created.
- Independent data marts, in contrast, are standalone systems built by drawing data directly from operational or external sources of data or both.
- Hybrid data marts can draw data from operational systems or data warehouses

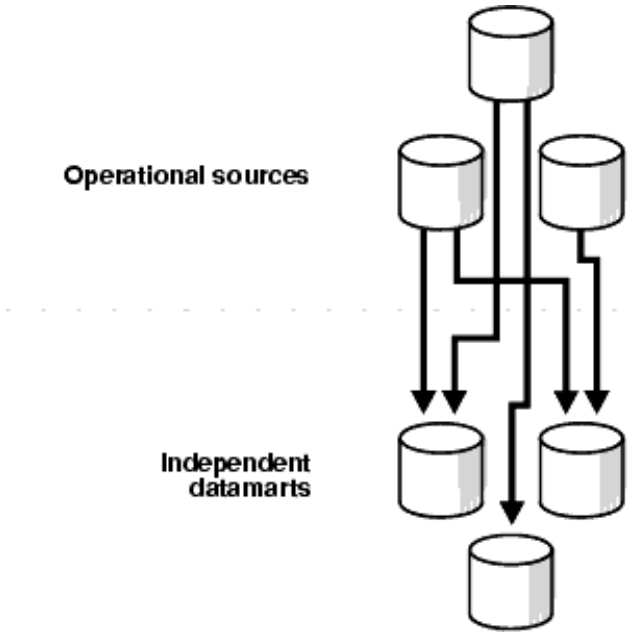
1. **Dependent Data Marts**

- 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.



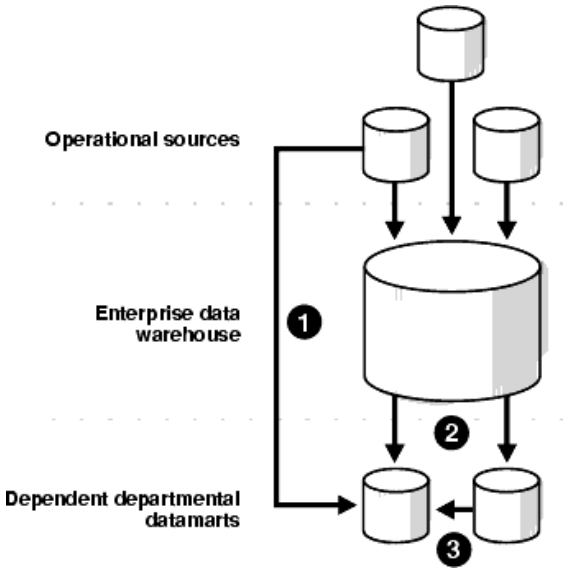
2. Independent Data Marts

- 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.

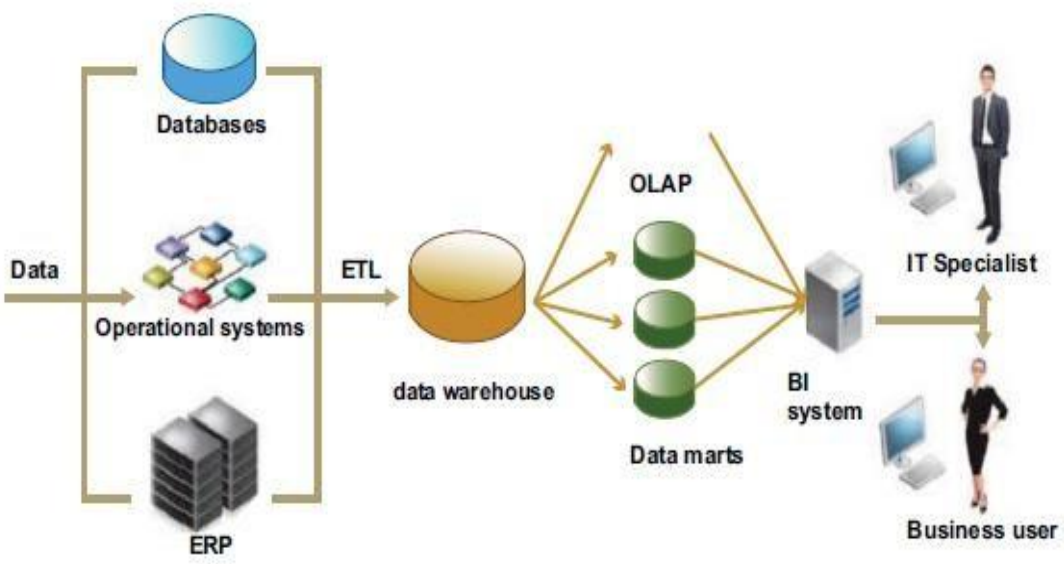


OR

Question-10

What is Business Intelligence? Explain the need for Business Intelligence today.

- While there are varying definitions for *BI*, Forrester defines it broadly as a "set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information that allows business users to make informed business decisions with real-time data that can put a company ahead of its competitors".
- In other words, the high-level goal of BI is to help a business user turn business-related data into actionable knowledge.



- BI traditionally focused on reports, dashboards, and answering predefined questions
- Today BI also includes a focus on deeper, exploratory, and interactive analyses of the data using *Business Analytics* such as data mining, predictive analytics, statistical analysis, and natural language processing solutions.
- BI systems evolved by adding layers of data staging to increase the accessibility of the business data to business users.
- Data from the operational systems and ERP were extracted, transformed into a more consumable form (e.g., column names labeled for human rather than computer consumption, errors corrected, duplication eliminated).
- Data from a warehouse were then loaded into OLAP cubes, as well as data marts stored in data warehouses.
- OLAP cubes facilitated the analysis of data over several dimensions.
- Data marts present a subset of the data in the warehouse, tailored to a specific line of business.
- Using Business Intelligence, the business user, with the help of an IT specialist who had set up the system for her, could now more easily access and analyze the data through a BI system.