CMR INSTITUTE OF TECHNLOGY

USN | | | | | | | | | |

Internal Assessment Test-1

Sub:	**Data Mining and Data Warehousing**					Sub Code:	**15CS561**	Branch :	**CSE/ISE**
Date:	7/03/19	Duration:	90 min's	Max Marks:	50	Sem/Sec:	VI A , B,C		OBE

Answer any FIVE FULL Questions

		MARKS	CO	RBT
1 (a)	Give the definition of data warehousing with a schematic diagram, explain the working of general data warehousing architecture.	[10]	CO1	L2
2 (a)	Distinguish between OLTP and OLAP	[10]	CO2	L2
3 (a)	Explain the operation of data cube with suitable example	[10]	CO2	L2
4 (a)	Explain various multidimensional data models	[10]	CO1	L2
5 (a)	Discuss the challenges that motivate the development of data mining	[5]	CO3	L2
(b)	Explain the difference between ROLAP and MOLAP	[5]	CO2	L2
6 (a)	Write short note on the following: i).Extraction ii).Cleaning iii).Transformation iv).Loading and Refreshing v)Meta data Repository	[10]	CO1	L1
7 (a)	Explain the role of Concept Hierarchies with suitable example	[5]	CO2	L2
(b)	Describe the Categorization of Measures with suitable example	[5]	CO2	L1

----------All The Best ----------

Solution for IAT- 1

Date of Exam : 07.03.2019

15CS561 - Data Mining and Data Warehousing

1. Give the definition of data warehousing with a schematic diagram, explain the working of general data warehousing architecture.

Definition of Data warehouse:

A data warehouse is a subject-oriented, integrated, time-variant, and non - volatile collection of data in support of management's decision making process"

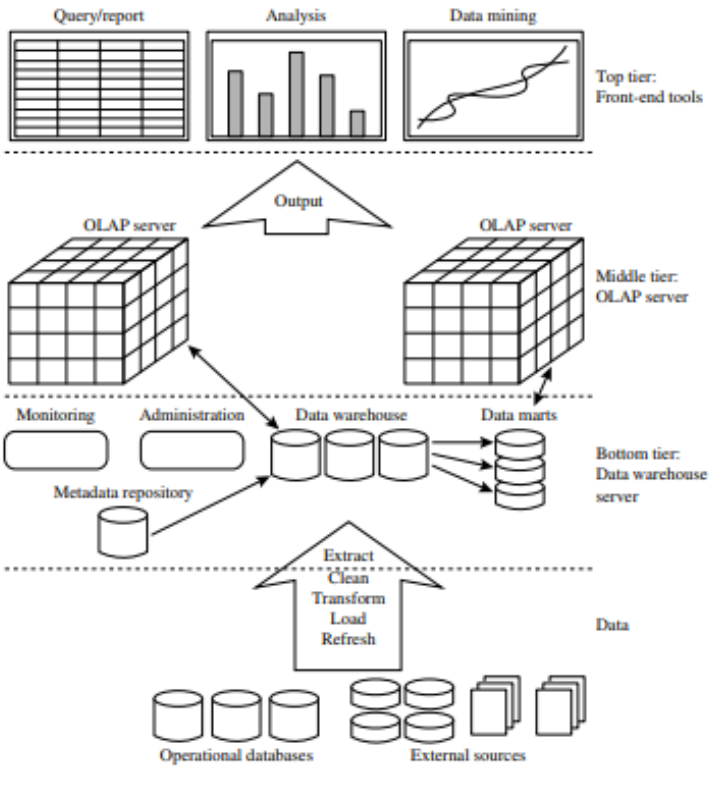


Figure 4.1 A three-tier data warehousing architecture.

Explanation for Three – tier architecture:

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

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. 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 (1) a relational OLAP (ROLAP) model (i.e., an extended relational DBMS that maps operations on multidimensional data to standard relational operations); or (2) a multidimensional OLAP (MOLAP) model (i.e., 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 (e.g., trend analysis, prediction, and so on).

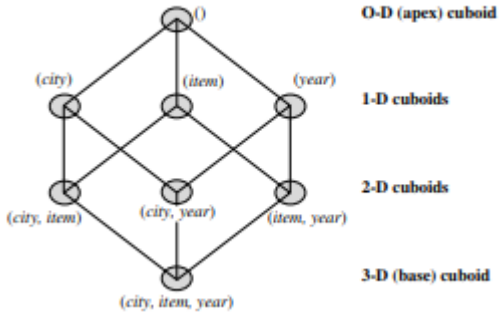
2. Distinguish between OLTP and OLAP

Table 4.1 Comparison of OLTP and OLAP Systems

Feature	*OLTP*	*OLAP*
Characteristic	operational processing	informational processing
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	historic, 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
Focus	data in	information out
Operations	index/hash on primary key	lots of scans
Number of records accessed	tens	millions
Number of users	thousands	hundreds
DB size	GB to high-order GB	$\geq$ TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time

3. Explain the operations of data cube with suitable example

A data cube allows data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts. In general terms, dimensions are the perspectives or entities with respect to which an organization wants to keep records. For example, sales data warehouse keeps records of the store's sales with respect to the dimensions time, item, branch, and location. These dimensions allow the store to keep track of things like monthly sales of items and the branches and locations at which the items were sold. Each dimension may have a table associated with it, called a dimension table, which further describes the dimension.



Operations are:

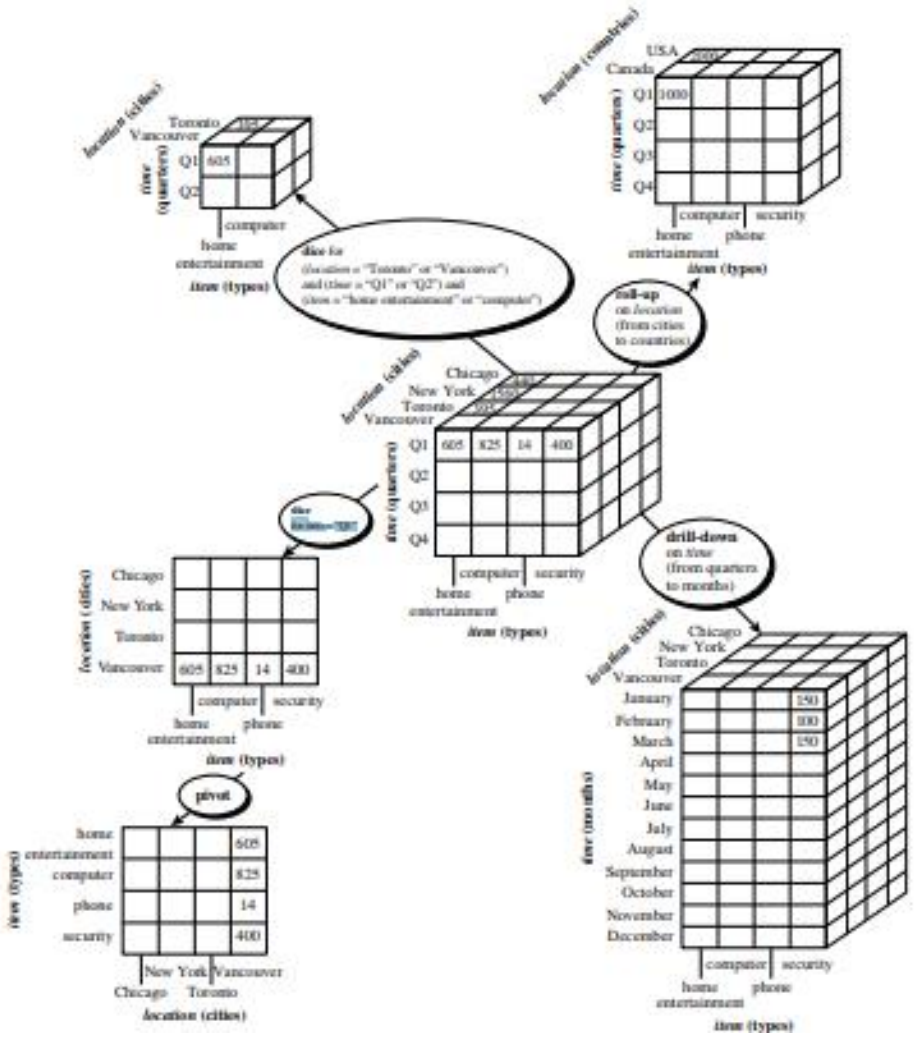
Roll-up : The roll-up operation performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction.

Drill down: Drill-down is the reverse of roll-up. It navigates from less detailed data to more detailed data. Drill-down can be realized by either stepping down a concept hierarchy for a dimension or introducing additional dimensions.

Slicing: The slice operation performs a selection on one dimension of the given cube, resulting in a subcube.

Dicing: The dice operation defines a subcube by performing a selection on two or more dimensions.

Pivoting: Pivot (or rotate) is a visualization operation that rotates the data axes in view to provide an alternative data presentation.

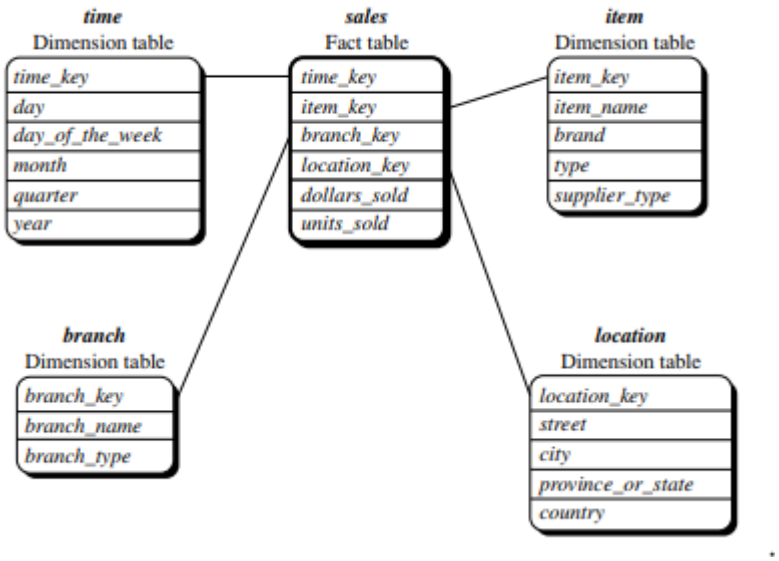


4. Explain various multidimensional data models.

Data warehouses and OLAP tools are based on a multidimensional data model. This model views data in the form of a data cube. A data cube allows data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.

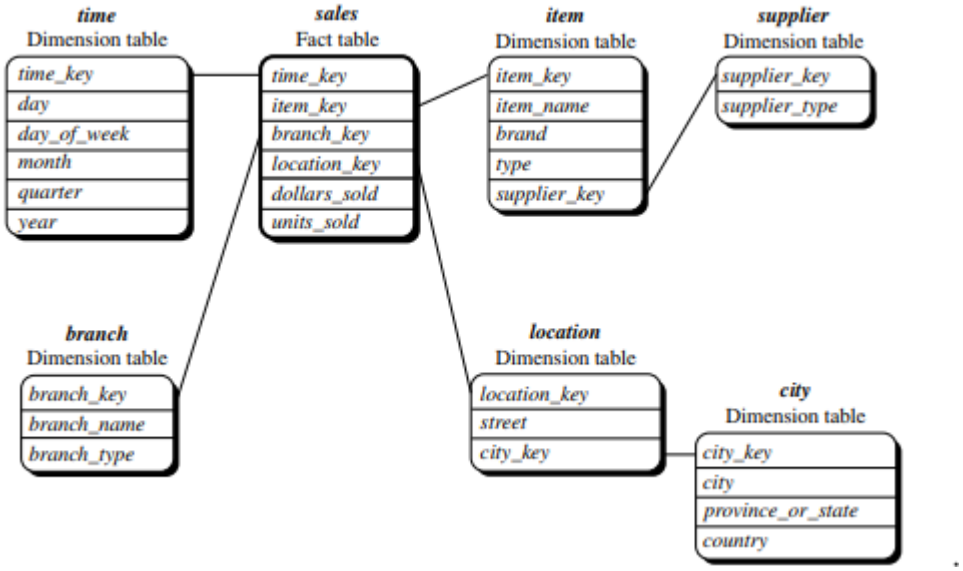
The most popular data model for a data warehouse is a multidimensional model, which can exist in the form of a star schema, a snowflake schema, or a fact constellation schema.

a) Star Schema



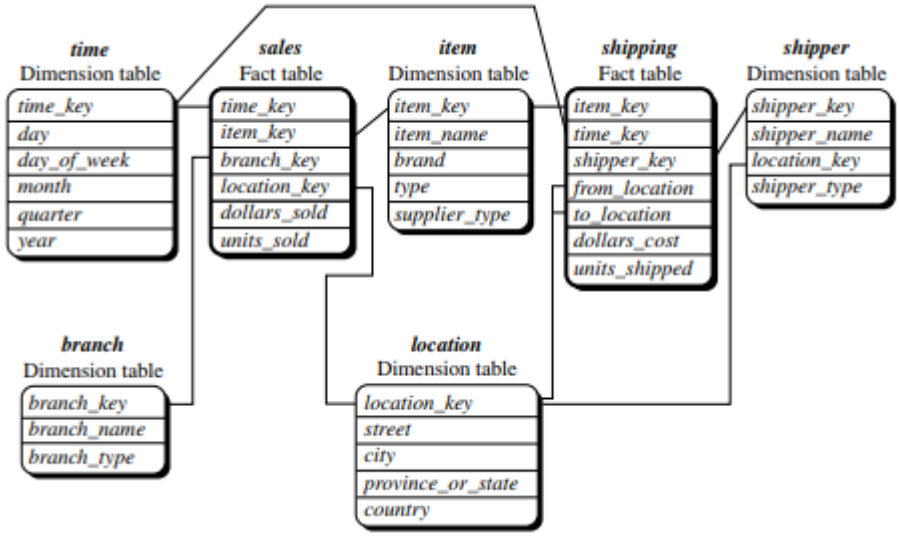
The most common modeling paradigm is the star schema, in which the data warehouse contains (1) a large central table (fact table) containing the bulk of the data, with no redundancy, and (2) a set of smaller attendant tables (dimension tables), one for each dimension. The schema graph resembles a starburst, with the dimension tables displayed in a radial pattern around the central fact table.

b) Snow flake scheme



The snowflake schema is a variant of the star schema model, where some dimension tables are normalized, thereby further splitting the data into additional tables. The resulting schema graph forms a shape similar to a snowflake. 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.

c) 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.

5.a) Discuss the challenges that motivate the development of data mining. (5 marks for any 5 points)

1. Handling complex types of data: Diverse applications generate a wide spectrum of new data types, from structured data such as relational and data warehouse data to semi-structured and unstructured data; from stable data repositories to dynamic data streams; from simple data objects to temporal data, biological sequences, sensor data, spatial data, hypertext data, multimedia data, software program code, Web data, and social network data.

2. ***Mining dynamic, networked, and global data repositories:*** Multiple sources of data are connected by the Internet and various kinds of networks, forming gigantic, distributed, and heterogeneous global information systems and networks.

3. ***Mining various and new kinds of knowledge:*** Data mining covers a wide spectrum of data analysis and knowledge discovery tasks, from data characterization and discrimination to association and correlation analysis, classification, regression, clustering, outlier analysis, sequence analysis, and trend and evolution analysis.

4. Mining knowledge in multidimensional space: When searching for knowledge in large data sets, we can explore the data in multidimensional space.

5. Data mining—an interdisciplinary effort: The power of data mining can be substantially enhanced by integrating new methods from multiple disciplines to mine data with natural language text, it makes sense to fuse data mining methods with methods of information retrieval and natural language processing.

6. Handling uncertainty, noise, or incompleteness of data: Data often contain noise, errors, exceptions, or uncertainty, or are incomplete. Errors and noise may confuse the data mining process, leading to the derivation of erroneous patterns.

7. ***Pattern evaluation and pattern- or constraint-guided mining:*** Not all the patterns generated by data mining processes are interesting. What makes a pattern interesting may vary from user to user.

5.b) Explain the difference between ROLAP and MOLAP

BASIS FOR COMPARISON	ROLAP	MOLAP
Full Form	ROLAP stands for Relational Online Analytical Processing.	MOLAP stands for Multidimensional Online Analytical Processing.
Storage & Fetched	Data is stored and fetched from the main data warehouse.	Data is Stored and fetched from the Proprietary database MDDBs.
Data Form	Data is stored in the form of relational tables.	Data is Stored in the large multidimensional array made of data cubes.
Data volumes	Large data volumes.	Limited summaries data is kept in MDDBs.
Technology	Uses Complex SQL queries to fetch data from the main warehouse.	MOLAP engine created a pre calculated and prefabricated data cubes for multidimensional data views. Sparse matrix technology is used to manage data sparsity.
View	ROLAP creates a multidimensional view of data dynamically.	MOLAP already stores the static multidimensional view of data in MDDBs.
Access	Slow access.	Faster access.

6.Write short note on the following:
i).Extraction
ii).Cleaning
iii).Transformation
iv).Loading and Refreshing
v)Meta data Repository

i) **Data extraction** is typically gathers data from multiple, heterogeneous, and external sources.

ii) **Data cleaning** detects errors in the data and rectifies them when possible.

iii) **Data transformation** process converts data from legacy or host format to warehouse format.

iv) **Loading and Refreshing:**

Load, which sorts, summarizes, consolidates, computes views, checks integrity, and builds indices and partitions.

Refresh, which propagates the updates from the data sources to the warehouse.

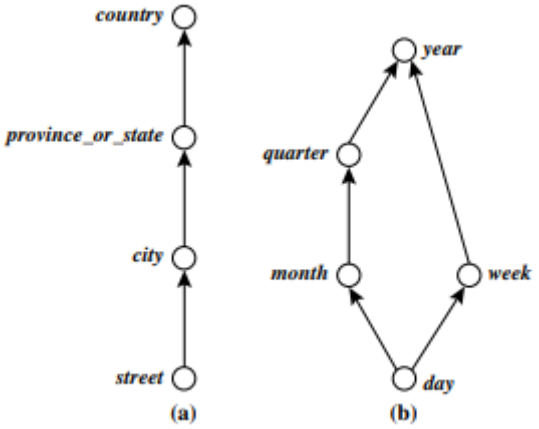
v) **Meta data Repository :**

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. A metadata repository should contain the following: A description of the data warehouse structure, which includes the warehouse schema, view, dimensions, hierarchies, and derived data definitions, as well as data mart locations and contents.

7. a) Explain the role of Concept Hierarchies with suitable example.

A concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts. Consider a concept hierarchy for the dimension location. City values for location include Vancouver, Toronto, New York, and Chicago. Each city, however, can be mapped to the province or state to which it belongs. For example, Vancouver can be mapped to British Columbia, and Chicago to Illinois. The provinces and states can in turn be mapped to the country (e.g., Canada or the United States) to which they belong. These mappings form a concept hierarchy for the dimension location, mapping a set of low-level

concepts (i.e., cities) to higher-level, more general concepts (i.e., countries). This concept hierarchy is illustrated in figure.



7.b) Describe the Categorization of Measures with suitable example.

A data cube measure is a numeric function that can be evaluated at each point in the data cube space. Measures can be organized into three categories—***distributive, algebraic, and holistic***—based on the kind of aggregate functions used.

Distributive: An aggregate function is distributive if it can be computed in a distributed manner as follows. Suppose the data are partitioned into n sets. We apply the function to each partition, resulting in n aggregate values. If the result derived by applying the function to the n aggregate values is the same as that derived by applying the function to the entire data set (without partitioning), the function can be computed in a distributed manner. For example, sum() can be computed for a data cube by first partitioning the cube into a set of subcubes, computing sum() for each subcube, and then summing up the counts obtained for each subcube. Hence, sum() is a distributive aggregate function. For the same reason, count(), min(), and max() are distributive aggregate functions. By treating the count value of each nonempty base cell as 1 by default, count() of any cell in a cube can be viewed as the sum of the count values of all of its corresponding child cells in its subcube. Thus, count() is distributive.

A measure is distributive if it is obtained by applying a distributive aggregate function. Distributive measures can be computed efficiently because of the way the computation can be partitioned.

Algebraic: An aggregate function is algebraic if it can be computed by an algebraic function with M arguments (where M is a bounded positive integer), each of which is obtained by applying a distributive aggregate function. For example, avg() (average) can be computed by sum()/count(), where both sum() and count() are distributive aggregate functions. Similarly, it can be shown that min N() and max N() (which find the N minimum and N maximum values, respectively, in a given set) and standard deviation() are algebraic aggregate functions. A measure is algebraic if it is obtained by applying an algebraic aggregate function.

Holistic: An aggregate function is holistic if there is no constant bound on the storage size needed to describe a subaggregate. That is, there does not exist an algebraic function with M arguments (where M is a constant) that characterizes the computation. Common examples of holistic functions include median(), mode(), and rank(). A measure is holistic if it is obtained by applying a holistic aggregate function.
