USN										

CMRIT — CMR INSTITUTE OF TECHNOLOGY, BENGALURU. ACCREDITED WITH A+ GRADE BY NAAC

Internal Assessment Test 1 – May 2022

Sub:	DATA MINING AND DATA WAREHOUSING				Sub Code:	18CS641	Branch:	ISE
Date:	10/05/2022	Duration:	90 min's	Max Marks: 50	Sem/Sec:	VI / A, B & C		OBE

	<u>Answer any FIVE FULL Questions</u>	MARKS	CO	RBT
1 (a)	Define Data warehouse. Explain its Key features.	[05]	CO1	L1
(b)	Differentiate OLTP with OLAP in terms of various criterions.	[05]	CO1	L2
2	Explain the 3-tier architecture of Data warehouse in detail with a neat diagram.	[10]	CO1	L2
3	Suppose that a data warehouse consists of the three dimensions time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit. a. Enumerate three classes of schemas that are popularly used for modelling data warehouses using Star Schema. b. Draw star and snowflake schema diagram for the above data warehouse. c. Starting with the base cuboid [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2010. d. To obtain the same list, write an SQL query assuming the data are stored in a relational database with the schema fee (day, month, year, doctor, hospital, patient, count, charge).	[10]	CO1	L3
4	Explain with suitable examples and diagrams the OLAP operations in multi-dimensional data Model.	[10]	CO1	L2
5 (a)	Write a short note on Compute Cube Operator and Curse of Dimensionality.	[05]	CO2	L1
(b)	Explain the concept of Materialization for the Selected Computation of Cuboids.	[05]	CO2	L2
6	Explain indexing OLAP Data: Bitmap Index and Join Index with an example.	[10]	CO2	L2

Faculty Signature CCI Signature HOD Signature

CMRIT

Internal Assessment Test 1 – May 2022
Scheme of Evaluation

Sub:	DATA MINING AND DATA WAREHOUSING					Sub Code:	18CS641	Branch:	ISE
Date:	10/05/2022	Duration:	90 min's	Max Marks:	50	Sem/Sec:	VI / A, B & C		OBE

Answer any FIVE FULL Questions — MARKS | CO | RBT

1 (a) Define Data warehouse. Explain its Key features. [05] CO1 L1

Scheme:- Definition + explanation of features= 2+3 M = 5M

Solution:-

Data warehousing provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions.

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

Features:

- Subject-Oriented: A data warehouse can be used to analyse a particular subject area.
- Integrated: A data warehouse integrates data from multiple data sources.
- Time-Variant: Historical data is kept in a data warehouse.
- Non-volatile: Once data is in the data warehouse, it will not change. So, historical data in a data warehouse should never be altered.

(b) Differentiate OLTP with OLAP in terms of various criterions. [05] CO1 L2

Scheme: Differences of OLAP & OLTP with atleast 10 criterions:- 5 Marks

Solution:-

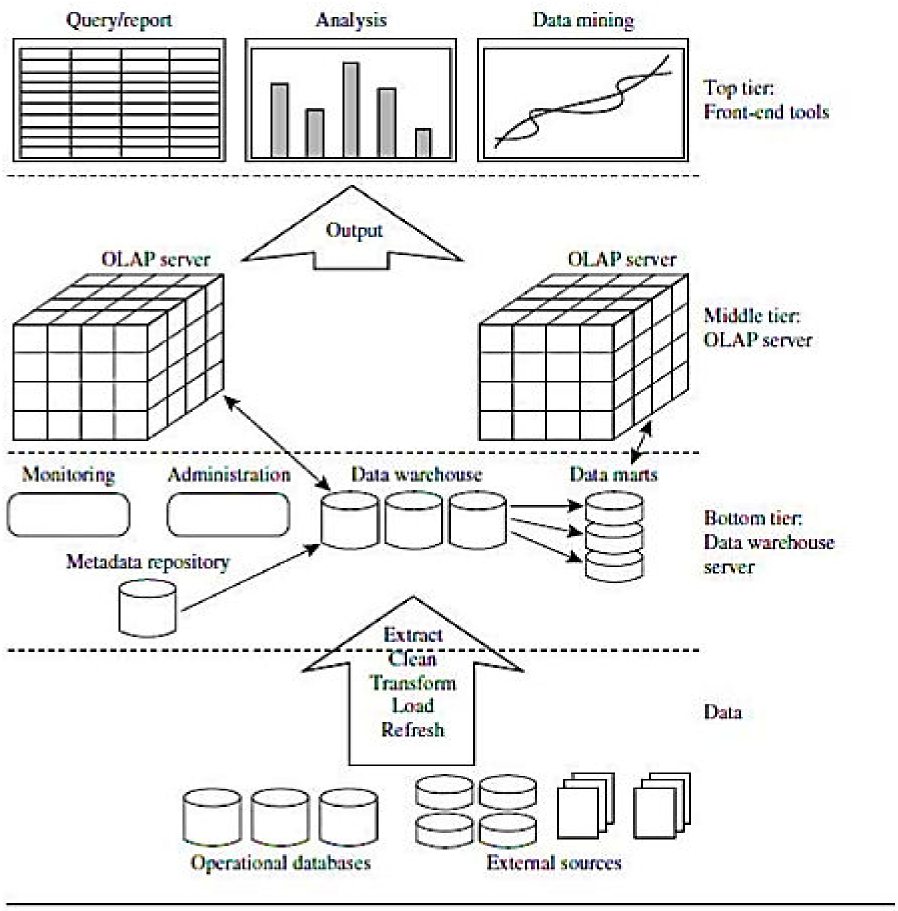
Feature	OLTP	OLAP
Characteristic	operational processing	informational processing
Orientation	transaction	analysis
User	clerk, DBA, database professional	knowledge worker (e.g., manager, executive, analyst)
Function	day-to-day operations	long-term informational requirements decision support
DB design	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	≥ TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time

<table>
<tr>
<td>2</td>
<td>

Explain the 3-tier architecture of Data warehouse in detail with a neat diagram.

Scheme:- Explanation of all the Tiers + Diagram = 7+3 =10 Marks

Solution:-



- The bottom tier is a warehouse database server that is almost always a relational database system.
- A gateway is supported by the underlying DBMS and allows client programs to generate SQL code to be executed at a server.
- The middle tier is an OLAP server that is typically implemented using either (1) a relational OLAP (ROLAP) model or (2) a multidimensional OLAP (MOLAP) model.
- The top tier is a front-end client layer, which contains query and reporting tools, analysis tools, and/or data mining tools.

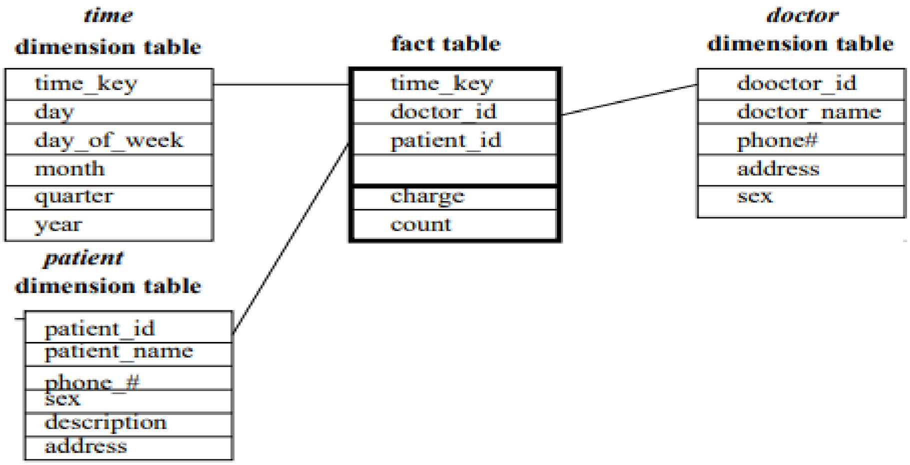
</td>
<td>[10]</td>
<td>CO1</td>
<td>L2</td>
</tr>
<tr>
<td>3</td>
<td>

Suppose that a data warehouse consists of the three dimensions time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit.

a. Enumerate three classes of schemas that are popularly used for modelling data warehouses using Star Schema.
Scheme:- Star Schema Definition – 1 Mark
Solution:- A fact table in the middle connected to a set of dimension tables.

b. Draw star and snowflake schema diagram for the above data warehouse.
Scheme:- Star and Snowflake Schema for Doctor Warehouse- 4 Marks
Solution:-

</td>
<td>[10]</td>
<td>CO1</td>
<td>L3</td>
</tr>
</table>



c. Starting with the base cuboid [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2010.

Scheme:- Defining OLAP Operations for cuboid – 3 Marks

Solution:- The operations to be performed are:

- Roll-up on time from day to year.
- Slice for time = 2010.
- Roll-up on patient from individual patient to all.

d. To obtain the same list, write an SQL query assuming the data are stored in a relational database with the schema fee (day, month, year, doctor, hospital, patient, count, charge).

Scheme:- Writing SQL Query – 2 Marks

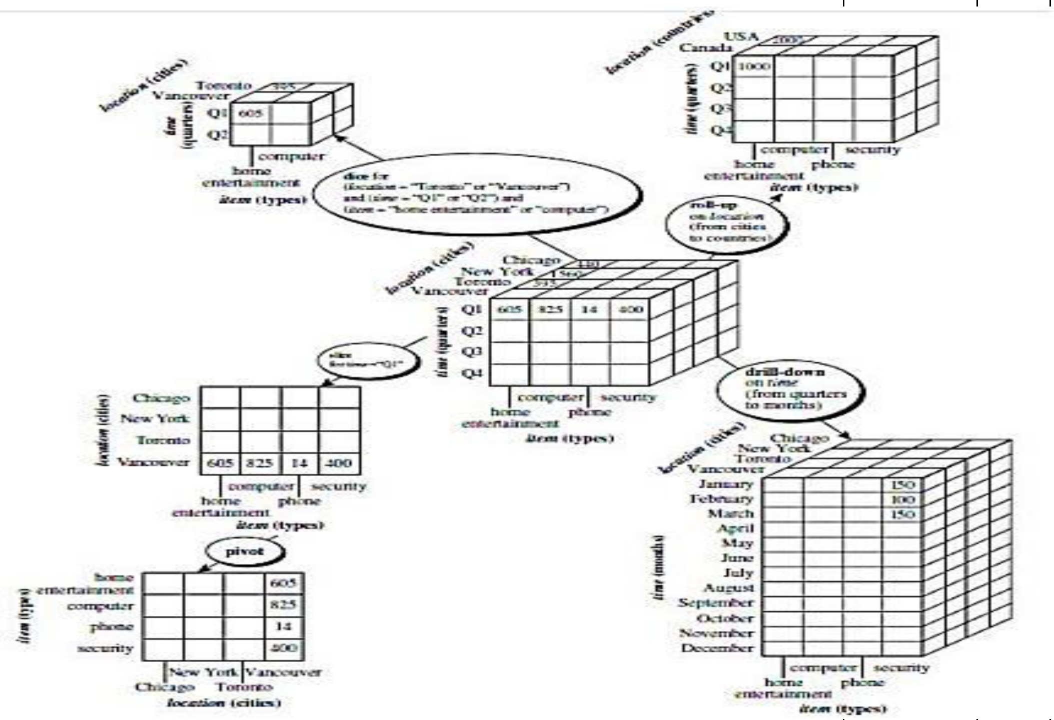
Solution:-

select doctor, SUM(charge) from fee where year = 2010 group by doctor

4	**Explain with suitable examples and diagrams the OLAP operations in multi-dimensional data Model.**	[10]	CO1	L2

Scheme:- Defining, explaining OLAP Operations with examples +Diagram–7+3 Marks =10 Marks

Solution:-



	• The **roll-up** operation also called as the drill-up operation performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction. • **Drill-down** can be realized by either stepping down a concept hierarchy for a dimension or introducing additional dimensions. • The **slice** operation performs a selection on one dimension of the given cube, resulting in a subcube and the **dice** operation defines a subcube by performing a selection on two or more dimensions. • **Pivot (also called rotate)** is a visualization operation that rotates the data axes in view to provide an alternative data presentation.			
5 (a)	**Write a short note on Compute Cube Operator and Curse of Dimensionality.** **Scheme:-** Compute Cube Operator + Curse Of Dimensionality = 2+3 Marks= 5 Marks **Solution:-** A **cube operator** on n dimensions is equivalent to a collection of group-by statements, one for each subset of the n dimensions Ex:- define cube sales_cube [city, item, year]: sum(sales in dollars) **Curse Of Dimensionality**:- How many cuboids in an n-dimensional cube with L levels? • If there were no hierarchies associated with each dimension, then the total number of cuboids for an n-dimensional data cube, as we have seen, is 2^n . However, in practice, many dimensions do have hierarchies. For example, time is usually explored not at only one conceptual level (e.g., year), but rather at multiple conceptual levels such as in the hierarchy "day < month < quarter < $\textit{Total number of cuboids} = \prod_{i=1}^{n}(L_i + 1),$	[05]	CO2	L1
(b)	**Explain the concept of Materialization for the Selected Computation of Cuboids.** **Scheme:-** Explanation of all types of Materialization : 1+1+3 Marks = 5 Marks **Solution:-** 1.No materialization: Do not pre-compute any of the "non base" cuboids. This leads to computing expensive multidimensional aggregates on-the-fly, which can be extremely slow. 2.Full materialization: Pre-compute all of the cuboids. The resulting lattice of computed cuboids is referred to as the full cube. This choice typically requires huge amounts of memory space in order to store all of the pre-computed cuboids. 3.Partial materialization: Selectively compute a proper subset of the whole set of possible cuboids. Alternatively, we may compute a subset of the cube, which contains only those cells that satisfy some user-specified criterion, such as where the tuple count of each cell is above some threshold The partial materialization of cuboids or subcubes should consider three factors: (1) identify the subset of cuboids or subcubes to materialize; (2) exploit the materialized cuboids or subcubes during query processing; and (3) efficiently update the materialized cuboids or subcubes during load and refresh.	[05]	CO2	L2
6	**Explain indexing OLAP Data: Bitmap Index and Join Index with an example.** **Scheme:-** Explanation of Bitmap Index and Join Index with an example each : 5+5 Marks = 10 Marks.	[10]	CO2	L2

Solution:-

- In the bitmap index for a given attribute, there is a distinct bit vector, Bv, for each value v in the attribute's domain.
- If the attribute has the value v for a given row in the data table, then the bit representing that value is set to 1 in the corresponding row of the bitmap index. All other bits for that row are set to 0.

Base table

RID	*item*	*city*
R1	H	V
R2	C	V
R3	P	V
R4	S	V
R5	H	T
R6	C	T
R7	P	T
R8	S	T

item bitmap index table

RID	H	C	P	S
R1	1	0	0	0
R2	0	1	0	0
R3	0	0	1	0
R4	0	0	0	1
R5	1	0	0	0
R6	0	1	0	0
R7	0	0	1	0
R8	0	0	0	1

city bitmap index table

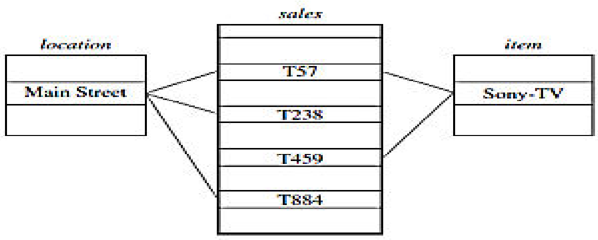
RID	V	T
R1	1	0
R2	1	0
R3	1	0
R4	1	0
R5	0	1
R6	0	1
R7	0	1
R8	0	1

Note: H for "home entertainment," C for "computer," P for "phone," S for "security," V for "Vancouver," T for "Toronto."

Indexing OLAP data using bitmap indices.

- The join indexing method gained popularity from its use in relational database query processing. Traditional indexing maps the value in a given column to a list of rows having that value.
- In contrast, join indexing registers the joinable rows of two relations from a relational database. For example, if two relations R(RID, A) and S(B, SID) join on the attributes A and B, then the join index record contains the pair (RID, SID), where RID and SID are record identifiers from the R and S relations, respectively.

.



5 Linkages between a *sales* fact table and *location* and *item* dimension tables.

Linkages between a *sales* fact table and *location* and *item* dimension tables.

Join index table for *location/sales*

location	*sales_key*
...	...
Main Street	T57
Main Street	T238
Main Street	T884
...	...

Join index table for *item/sales*

item	*sales_key*
...	...
Sony-TV	T57
Sony-TV	T459
...	...

Join index table linking *location* and *item* to *sales*

location	*item*	*sales_key*
...	...	...
Main Street	Sony-TV	T57
...	...	...

Faculty Signature	CCI Signature	HOD Signature