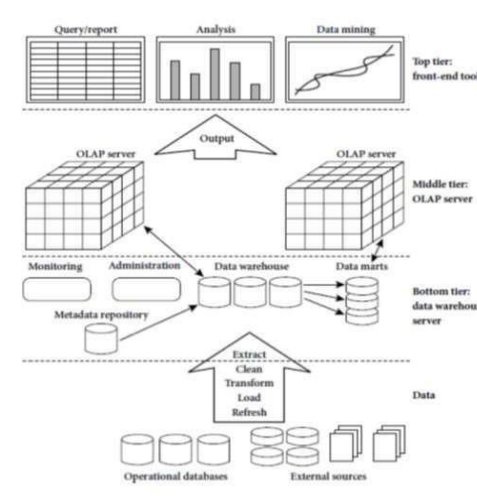


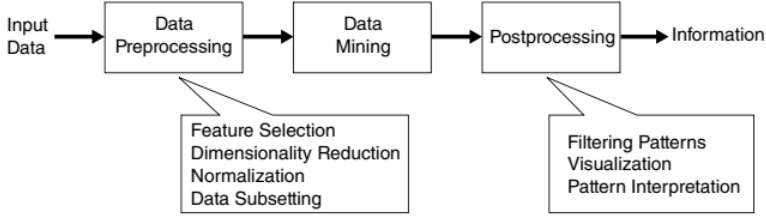
USN

### Internal Assessment Test 1 – June-2024

Sub:	<b>DATA MINING AND DATA WAREHOUSING</b>	Sub Code:	<b>21IS643</b>	Branch:	<b>ISE</b>
Date:	<b>03/06/2024</b>	Duration:	<b>90 min</b>	Max Marks:	<b>50</b>
				Sem/Sec:	<b>VI / C</b>
					<b>OBE</b>

#### Answer any FIVE FULL Questions

		MARKS	CO	RBT																																																						
1 (a)	Give the different views on the definitions of Data warehouse. With its Key features. Sol : Definitions (Subject , Time, Nonvolatile and integrated –explain the keywords)	[05]	CO1	L1																																																						
(b)	List the difference between OLTP from OLAP in terms of various functionalities. Sol :Any 5 points <b>Table 4.1</b> Comparison of OLTP and OLAP Systems	[05]	CO1	L1																																																						
<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="width: 25%;">Feature</th> <th style="width: 35%;">OLTP</th> <th style="width: 35%;">OLAP</th> </tr> </thead> <tbody> <tr> <td>Characteristic</td> <td>operational processing</td> <td>informational processing</td> </tr> <tr> <td>Orientation</td> <td>transaction</td> <td>analysis</td> </tr> <tr> <td>User</td> <td>clerk, DBA, database professional</td> <td>knowledge worker (e.g., manager, executive, analyst)</td> </tr> <tr> <td>Function</td> <td>day-to-day operations</td> <td>long-term informational requirements decision support</td> </tr> <tr> <td>DB design</td> <td>ER-based, application-oriented</td> <td>star/snowflake, subject-oriented</td> </tr> <tr> <td>Data</td> <td>current, guaranteed up-to-date</td> <td>historic, accuracy maintained over time</td> </tr> <tr> <td>Summarization</td> <td>primitive, highly detailed</td> <td>summarized, consolidated</td> </tr> <tr> <td>View</td> <td>detailed, flat relational</td> <td>summarized, multidimensional</td> </tr> <tr> <td>Unit of work</td> <td>short, simple transaction</td> <td>complex query</td> </tr> <tr> <td>Access</td> <td>read/write</td> <td>mostly read</td> </tr> <tr> <td>Focus</td> <td>data in</td> <td>information out</td> </tr> <tr> <td>Operations</td> <td>index/hash on primary key</td> <td>lots of scans</td> </tr> <tr> <td>Number of records accessed</td> <td>tens</td> <td>millions</td> </tr> <tr> <td>Number of users</td> <td>thousands</td> <td>hundreds</td> </tr> <tr> <td>DB size</td> <td>GB to high-order GB</td> <td>≥ TB</td> </tr> <tr> <td>Priority</td> <td>high performance, high availability</td> <td>high flexibility, end-user autonomy</td> </tr> <tr> <td>Metric</td> <td>transaction throughput</td> <td>query throughput, response time</td> </tr> </tbody> </table>					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
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																																																								
2 a)	Explain the three-tier architecture of the Data warehouse in detail with a neat diagram. Sol : <b>Sol : Diagram + Explanation</b>	[06]	CO1	L2																																																						
 <p style="text-align: center;"><b>A Three Tier Data Warehouse Architecture:</b></p>																																																										

b)	<p>Distinguish between ROLAP,MOLAP,HOLAP Sol:</p> <table border="1" data-bbox="188 118 770 562"> <thead> <tr> <th>Property</th> <th>MOLAP</th> <th>ROLAP</th> </tr> </thead> <tbody> <tr> <td>Data structure</td> <td>Multidimensional database using sparse arrays</td> <td>Relational tables (each cell is a row)</td> </tr> <tr> <td>Disk space</td> <td>Separate database for data cube; large for large data cubes</td> <td>May not require any space other than that available in the data warehouse</td> </tr> <tr> <td>Retrieval</td> <td>Fast(pre-computed)</td> <td>Slow(computes on-the-fly)</td> </tr> <tr> <td>Scalability</td> <td>Limited (cubes can be very large)</td> <td>Excellent</td> </tr> <tr> <td>Best suited for</td> <td>Inexperienced users, limited set of queries</td> <td>Experienced users, queries change frequently</td> </tr> <tr> <td>DBMS facilities</td> <td>Usually weak</td> <td>Usually very strong</td> </tr> </tbody> </table>	Property	MOLAP	ROLAP	Data structure	Multidimensional database using sparse arrays	Relational tables (each cell is a row)	Disk space	Separate database for data cube; large for large data cubes	May not require any space other than that available in the data warehouse	Retrieval	Fast(pre-computed)	Slow(computes on-the-fly)	Scalability	Limited (cubes can be very large)	Excellent	Best suited for	Inexperienced users, limited set of queries	Experienced users, queries change frequently	DBMS facilities	Usually weak	Usually very strong	[04]	CO1	L2
Property	MOLAP	ROLAP																							
Data structure	Multidimensional database using sparse arrays	Relational tables (each cell is a row)																							
Disk space	Separate database for data cube; large for large data cubes	May not require any space other than that available in the data warehouse																							
Retrieval	Fast(pre-computed)	Slow(computes on-the-fly)																							
Scalability	Limited (cubes can be very large)	Excellent																							
Best suited for	Inexperienced users, limited set of queries	Experienced users, queries change frequently																							
DBMS facilities	Usually weak	Usually very strong																							
3	<p>Explain OLAP operations and diagrams for Retail Sales in a multi-dimensional data Model Sol : Explain the ROLL UP, ROLL DOWN,PIVOT,SLICE and DICE operations +diagram</p>	[10]	CO1	L2																					
4	<p>Explain the significance of data mining, its challenges, and the process of KDD with a neat diagram Sol : Data Mining defn, challenges and diagram of KDD with explanation</p> 	[10]	CO2	L2																					
5 (a)	<p>Apply the proximity measure for the following vectors X and Y, and calculate the cosine similarity. <math>X = \{3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0\}</math>, <math>Y = \{1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2\}</math> Sol : <math>x = (3, 2, 0, 5, 0, 0, 0, 2, 0, 0)</math> <math>y = (1, 0, 0, 0, 0, 0, 0, 1, 0, 2)</math>  <math>x \cdot y = 3 * 1 + 2 * 0 + 0 * 0 + 5 * 0 + 0 * 0 + 0 * 0 + 0 * 0 + 2 * 1 + 0 * 0 + 0 * 2 = 5</math> <math>\ x\  = \sqrt{3 * 3 + 2 * 2 + 0 * 0 + 5 * 5 + 0 * 0 + 0 * 0 + 0 * 0 + 2 * 2 + 0 * 0 + 0 * 0} = 6.48</math> <math>\ y\  = \sqrt{1 * 1 + 0 * 0 + 0 * 0 + 0 * 0 + 0 * 0 + 0 * 0 + 0 * 0 + 1 * 1 + 0 * 0 + 2 * 2} = 2.24</math> <math>\cos(x, y) = 0.31</math></p>	[04]	CO2	L3																					
(b)	<p>Apply the simple matching and Jaccard coefficient for the given problem to compute the similarity <math>x = \{1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\}</math> <math>y = \{0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1\}</math> Sol : <math>SMC = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}} = \frac{0 + 7}{2 + 1 + 0 + 7} = 0.7</math>  <math>J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} = \frac{0}{2 + 1 + 0} = 0</math> <math>f_{01} = 2</math> the number of attributes where x was 0 and y was 1 <math>f_{10} = 1</math> the number of attributes where x was 1 and y was 0 <math>f_{00} = 7</math> the number of attributes where x was 0 and y was 0 <math>f_{11} = 0</math> the number of attributes where x was 1 and y was 1</p>	[06]	CO2	L3																					

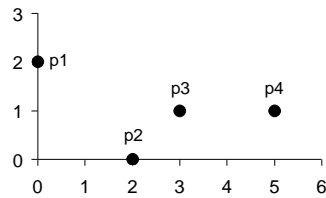
6

Compute the Euclidean distance and its importance in calculating the proximity measure  
For the given problem.

[10]

CO2

L3



$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2}$$

Sol :

	<b>p1</b>	<b>p2</b>	<b>p3</b>	<b>p4</b>
<b>p1</b>	0	2.828	3.162	5.099
<b>p2</b>	2.828	0	1.414	3.162
<b>p3</b>	3.162	1.414	0	2
<b>p4</b>	5.099	3.162	2	0

Faculty Signature

CCI Signature

HOD Signature