USN

Scheme and Solution of Internal Assessment Test 1 – Apr. 2023

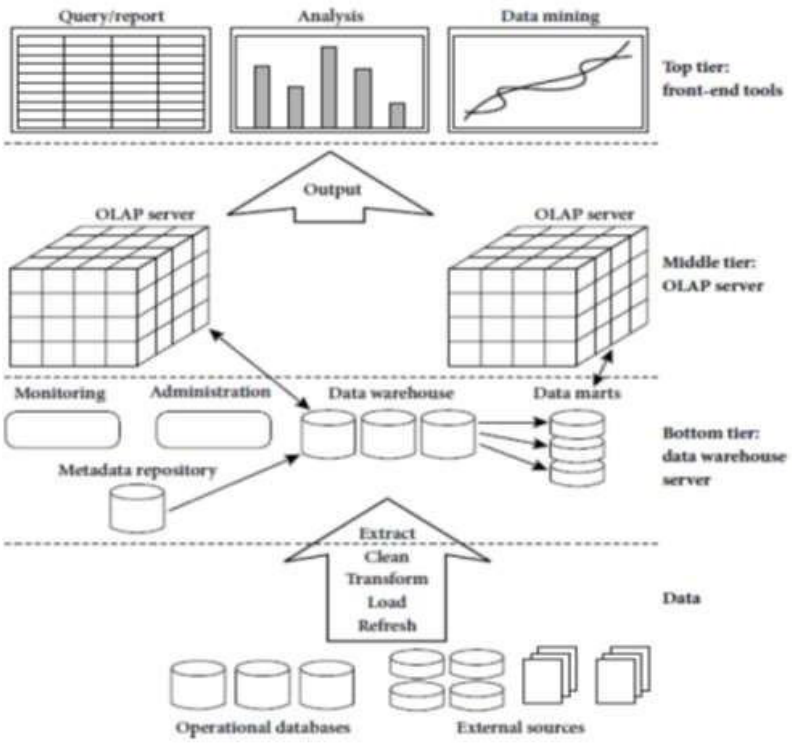
Sub:	**DATA MINING AND DATA WAREHOUSING**			Sub Code:	**18CS641**	Branch:	**CS**
Date:	25/04/2023	Duration:	90 mins	Max Marks:	50	Sem / Sec:	6 /A,B,C	OBE

	Answer any FIVE FULL Questions	MARKS	CO	RBT
1	a) Define data warehouse and explain four key features of data warehouse compare to other data repository systems Ans: A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process. **Key features:** 1. **Subject-Oriented:** A data warehouse can be used to analyze a particular subject area. For example, "sales" can be a particular subject. 2. **Integrated**: A data warehouse integrates data from multiple data sources. For example, source A and source B may have different ways of identifying a product, but in a data warehouse, there will be only a single way of identifying a product. 3. **Time-Variant**: Historical data is kept in a data warehouse. For example, one can retrieve data from 3 months, 6 months, 12 months, or even older data from a data warehouse. This contrasts with a transactions system, where often only the	[5+5]	CO1	L1

most recent data is kept. For example, a transaction system may hold the most recent address of a customer, where a data warehouse can hold all addresses associated with a customer.

4. **Non-volatile**: Once data is in the data warehouse, it will not change. So, historical data in a data warehouse should never be altered.

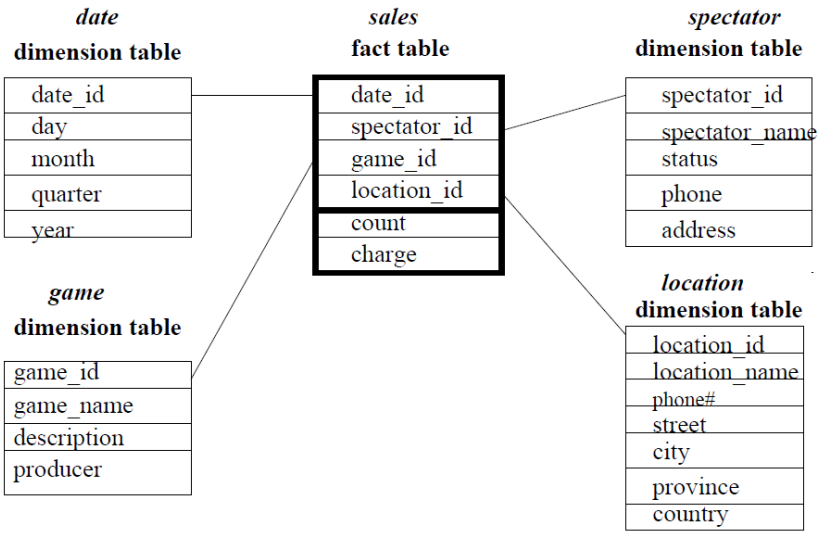
b) Explain the data warehouse multi- tier architecture with neat diagram



A Three Tier Data Warehouse Architecture:

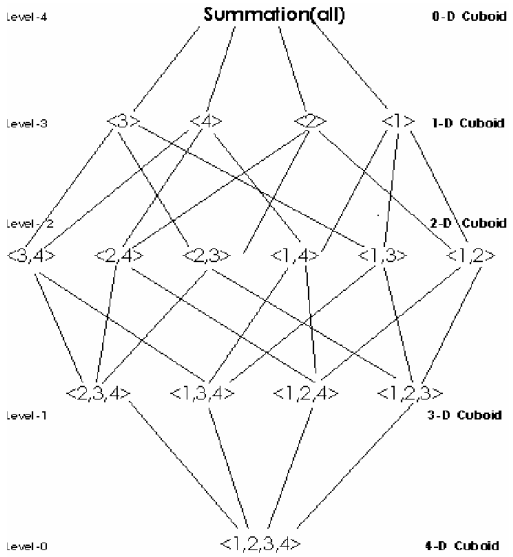
2	Explanation of each component must be written.
	
	a) Specify the following as OLTP or OLAP operations
	
	i. Retrieving money from ATM Machine- **OLTP**
	
	ii. Generating sales reports for a retail store that show sales by product, region, and time period **-OLTP**
	
	iii. Booking a flight ticket in an airline reservation system-**OLTP**
	
	iv. Analyzing website traffic data to identify trends in user behavior-**OLAP**
	
	v. Identifying top-selling products and analyzing their performance in different regions **-OLAP**
	
	b) Define concept hierarchies. Explain how concept hierarchy helps to analyze the sales data for the below given scenario?
	
	Consider a retail store that wants to analyze sales data for its electronics department. The store might create a concept hierarchy for the category "TVs" as follows:
	• TVs
	  • Brand
	  • Samsung
	  • LG
	  • Sony
	• Screen Size
	  • 32 inches and under
	  • 33 to 45 inches
	  • 46 to 55 inches
	  • 56 inches and above
	• Type
	  • LED
	  • OLED
	  • QLED
	
	Answer: Concept hierarchies are structures that organize concepts or ideas into a hierarchy or tree-like structure based on their relationships to one another. In a concept hierarchy, more general or abstract concepts are placed at the top of the hierarchy, while more specific or concrete concepts are placed further down the hierarchy.
		[5+5]	CO1	L2

	In this hierarchy, the top level is "TVs," and it has three subcategories: "Brand," "Screen Size," and "Type." Each of these subcategories is further broken down into more specific subcategories, such as "Samsung," "32 inches and under," and "LED." Using this concept hierarchy, the store can analyze sales data for TVs at different levels of abstraction. For example, the store could see that Samsung TVs are the most popular brand overall, or it could drill down to see that 55-inch and above OLED TVs from Samsung are selling particularly well. This kind of analysis can help the store make informed decisions about which TVs to stock, how to price them, and how to promote them to customers.			
3	Suppose that a data warehouse consists of the four dimensions, date, spectator, location, and game, and the two measures, count and charge, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate. (a) Draw a star schema diagram for the data warehouse. (5) (b) Create a lattice of cuboid using the four dimensions (3) (c)Starting with the base cuboid [date, spectator, location, game], what specific OLAP operations should one perform in order to list the total charge paid by student spectators at IPL_Chinnaswamy in 2022? (2) **Answer:** **a)Star Scheme:**	[5+3+2]	CO1	L3



b)Data cube using 4 dimensions.

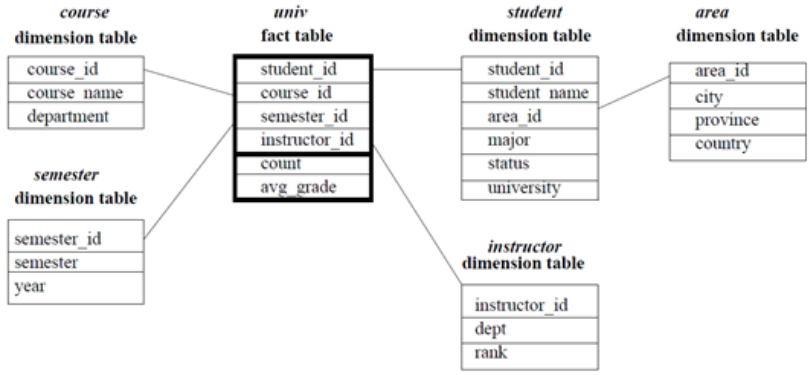
[1-date,2- spectator,3-location,4-game]

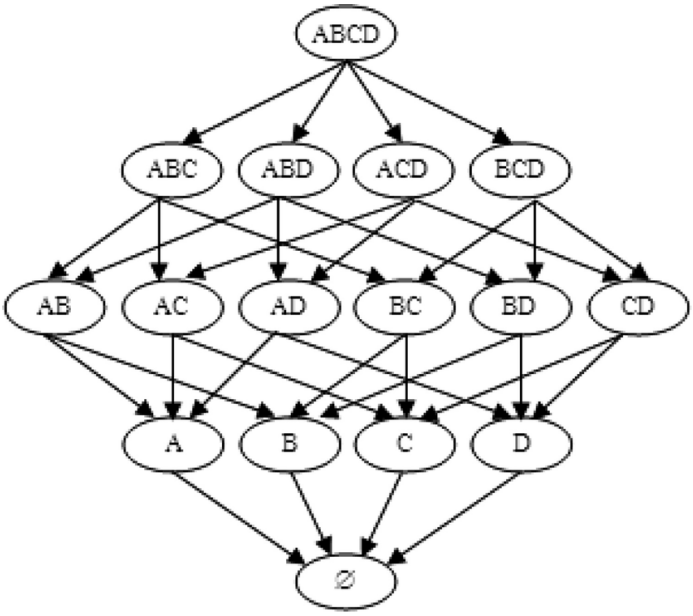


c) The specific OLAP operations to be performed are:

• Roll-up on date from date id to year.

	• Roll-up on game from game id to all. • Roll-up on location from location id to location name. • Roll-up on spectator from spectator id to status. • Dice with status="students", location name="IPL_Chinnaswamy", and year = 2022.			

4	 Figure -1 a) What data warehouse schema is represented in the Figure-1? b) Create a lattice of cuboid using [*student, course, semester, instructor*] dimensions. c)Starting with the base cuboid [*student, course, semester, instructor*], what specific *OLAP operations*(e.g., roll-up from *semester* to *year*) should one perform in order to list the average grade of *CS* courses for each student. (d) If each dimension has five levels (including all), such as "*student* < *major* < *status* < *university*< all", how many cuboids will this cube contain (including the base and apex cuboids)? e) Suppose that a data warehouse contains 20 dimensions, each with about five levels of granularity. Users are mainly interested in four particular dimensions, each having three frequently accessed levels for rolling up and drilling down. How would you design a data cube structure to support this preference efficiently? Answers: a) Snowflake schema b) Lattice of cuboid: [A-Student,B-Course, C-Semester,D-Instructor]	[10]	CO1	L3



c) Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations

(e.g., roll-up from semester to year) should one perform in order to list the average grade of CS courses for each Big-University student.

The specific OLAP operations to be performed are:

• Roll-up on course from course id to department.

• Roll-up on semester from semester id to all.

• Slice for course="CS" .

d) This cube will contain 5^4 = 625 cuboids.

e) The design of the data cube structure should focus on including only the necessary dimensions and levels, pre-aggregating data for frequently accessed levels, and creating indexes to optimize query performance. By doing so, we can efficiently support the user's preference for rolling up and drilling down on the four particular dimensions.

5	Describe the following terms: 1. Metadata repository 2. Data marts 3. Measures 4. Curse of dimensionality 5. ETL operations Answer: 1. Meta data Repository: A metadata repository is a centralized database or repository that stores metadata, which is data about data. The metadata in a repository may describe the structure, format, content, and context of various types of data and information resources, such as databases, data warehouses, documents, applications, and processes. 2. Data marts A data mart is a subset of a larger data warehouse that is designed to serve a specific business unit, department, or functional area within an organization. Data marts are typically smaller and more focused than a data warehouse, containing data that is specific to the needs of a particular group of users. The primary purpose of a data mart is to provide quick and easy access to relevant data for a specific business area or department. By focusing on a specific area of the business, data marts can provide more relevant and accurate information than a larger data warehouse, which may contain a lot of extraneous data that is not relevant to specific users. 3. Measures a measure is a numeric value that represents a quantifiable aspect of a business process or activity. Measures are typically used in conjunction with dimensions to create multidimensional models or data cubes that enable users to analyze data from different perspectives.	[10]	CO1	L2

	Examples of measures include sales revenue, profit, customer count, inventory level, and order volume. Measures are typically aggregated or summarized at different levels of granularity based on the dimensions of the data cube. For example, sales revenue may be aggregated by product, region, time period, or some combination of these dimensions. 4. Curse of dimensionality The curse of dimensionality is a phenomenon that can lead to significant challenges when working with high-dimensional data, including overfitting, underfitting, and sparsity. It is important to carefully consider the number of dimensions in a dataset when analyzing or modeling the data, and to use techniques such as dimensionality reduction to mitigate the effects of the curse of dimensionality. 5. ETL operations ETL (Extract, Transform, Load) is a process in data integration that involves extracting data from multiple sources, transforming it to fit the destination schema, and loading it into a target system, such as a data warehouse. ETL operations are commonly used in data integration, data migration, and data warehousing projects.			
6	a) Define datamining. Explain the process of knowledge discovery in databases (KDD) Answer: Data mining is the process of discovering hidden patterns, trends, and insights from large datasets, using statistical and machine learning techniques. It involves extracting useful and relevant information from data and transforming it into an understandable structure for further analysis. The process of knowledge discovery in databases (KDD) is a broader process that includes data mining as one of its key steps. KDD involves the entire process of discovering knowledge from data, including data selection, preprocessing, cleaning, transformation, data mining, and interpretation of the results. The KDD process typically involves the following steps:	[5+5]	CO1	L2

	Data Selection: In this step, the relevant data is identified and selected from various sources. The data can be structured or unstructured, and it can be obtained from databases, data warehouses, or other sources. Preprocessing: Once the data is selected, it is preprocessed to clean and transform it into a format suitable for data mining. This step includes tasks such as data cleaning, data transformation, and data reduction. Data Mining: This step involves applying statistical and machine learning algorithms to identify hidden patterns, relationships, and trends in the data. Common data mining techniques include clustering, classification, regression, and association rule mining. Interpretation/Evaluation: The results of the data mining process are evaluated and interpreted to identify meaningful patterns and insights. The interpretation of the results is usually done by domain experts, who can use the insights to make informed decisions. Application: The knowledge discovered in the previous steps is applied to solve real-world problems. The application of the knowledge can include making predictions, identifying opportunities, and making informed decisions. b) Discuss the challenges in data mining technologies Answer: Data mining technologies are essential for extracting useful insights and knowledge from large datasets. However, there are several challenges associated with data mining technologies, which can impact the quality and usefulness of the insights generated. Here are some of the key challenges in data mining technologies: Data Quality: The quality of the data used in data mining is critical to the accuracy and usefulness of the results. Poor data quality, including missing values, inconsistencies, and errors, can lead to inaccurate or misleading insights.			

	Data Volume: As the volume of data grows, the complexity and difficulty of analyzing it also increase. Large datasets can be challenging to analyze, as they require significant computational resources and specialized tools. Data Complexity: Data can be complex, and data mining techniques may not be able to identify all the relevant patterns and insights. Unstructured data, such as text or images, can be particularly challenging to analyze. Data Privacy and Security: Data mining technologies often require access to sensitive data, such as personal information, which can raise privacy and security concerns. The misuse of this data can lead to legal and ethical issues. Bias and Discrimination: Data mining algorithms can be biased, leading to unfair or discriminatory results. This can occur if the data used in the analysis is biased or if the algorithm itself is biased. Interpretability: Some data mining algorithms, such as deep learning, can be difficult to interpret, making it challenging to understand how the results were generated. This can make it difficult to validate the results and can undermine trust in the technology.			