DATA WAREHOUSING
(CSEN 3237)

Time Allotted : 2½ hrs **Full Marks : 60**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 4 (four) from Group B to E, taking one from each group.

Candidates are required to give answer in their own words as far as practicable.

Group – A

1. Answer any twelve: **12 × 1 = 12**

Choose the correct alternative for the following

(i) Which of the following is a characteristic of a data warehouse?
(a) Real-time data processing (b) Highly normalized data structure
(c) Subject-oriented data storage (d) Transactional processing system.

(ii) What is the main purpose of a data mart?
(a) To store historical data for reporting and analysis
(b) To integrate data from multiple sources into a centralized repository
(c) To provide a subset of data for a specific department or user group
(d) To perform online transaction processing.

(iii) Which architectural type does ROLAP represent in data warehousing?
(a) Relational (b) Multidimensional (c) Hybrid (d) Object-oriented.

(iv) A data warehouse is _____.
(a) updated by end users.
(b) contains numerous naming conventions and formats
(c) organized around important subject areas.
(d) contains only current data

(v) Transient data is _____.
(a) data in which changes to existing records cause the previous version of the records to be eliminated
(b) data in which changes to existing records do not cause the previous version of the records to be eliminated
(c) data that are never altered or deleted once they have been added
(d) data that are never deleted once they have been added

(vi) What is the purpose of partitioning in a data warehouse?
(a) To optimize data loading processes
(b) To increase data redundancy
(c) To improve query performance and manageability
(d) To enforce data integrity constraints.

(vii) The extract process is _____.
(a) capturing all of the data contained in various operational systems
(b) capturing a subset of the data contained in various operational systems
(c) capturing all of the data contained in various decision support systems
(d) capturing a subset of the data contained in various decision support systems

(viii) Which transformation type is used to identify and resolve duplicate records in data cleansing processes?
(a) Aggregation (b) Sorting (c) Deduplication (d) Filtering.

(ix) Data transformation includes _____.
(a) a process to change data from a detailed level to a summary level
(b) a process to change data from a summary level to a detailed level
(c) joining data from one source into various sources of data
(d) separating data from one source into various sources of data

(x) Business Intelligence and data warehousing is used for _____.
(a) Forecasting. (b) Data Mining
(c) Analysis of large volumes of product sales data. (d) All of the above.

Fill in the blanks with correct word

(xi) The _____ dimension represents time-related attributes such as day, month, quarter, and year.

(xii) Dimensionality reduction reduces the data set size by removing _____.

(xiii) _____ is a process of converting logical data model into physical database design structures such as tables, indexes, and views.

(xiv) _____ is a technique used to optimize query performance by limiting the amount of data scanned during query execution.

(xv) _____ is a good alternative to the star schema.

Group - B

2. (a) Suppose a data warehouse comprises four dimensions: employee, department, time, and location, along with two measures: Salary and Bonus. The attributes for each dimension and measures are as follows:
-**Dimension-Employee**:
EmployeeID - number(10) Primary Key, EmployeeName - varchar2(50) not null, Gender - varchar2(1), Position - varchar2(50), DepartmentID - number(10) Foreign Key.
-**Dimension-Department**:
DepartmentID - number(10) Primary Key, DepartmentName - varchar2(50) notnull, ManagerName - varchar2(50), LocationID - number(10) Foreign Key.
-**Dimension-Time**:
TimeKey - number(10) Primary Key, Date - date not null, Month - number(2) not null, Quarter - number(1) not null, Year - number(4) not null.
-**Dimension-Location**:
LocationID - number(10) Primary Key, City - varchar2(50), State - varchar2(50), Country - varchar2(50).
-**Measures**: Salary - numeric, Bonus - numeric.
(i) Enumerate the STAR schema that is popularly used for modelling data warehouses. *[(CO4)(Analyse/HOCQ)]*
(ii) Draw a STAR schema diagram for the provided data warehouse. *[(CO4)(Analyse/HOCQ)]*

(b) (i) What is slice operation?
(ii) What is dice operation?
(iii) What is pivot operation? *[(CO1)(Learn and understand/LOCQ)]*

2 + 4 + (2 + 2 + 2) = 12

3. (a) List out the OLAP operations in multidimensional data model? *[CSEN3237.1(Learn and understand/LOCQ)]*

(b) Answer the following:
(i) What is roll-up operation?
(ii) What is drill-down operation?
(iii) What is slice operation?
(iv) What is dice operation?
(v) What is pivot operation? *[CSEN3237.1(Learn and understand/LOCQ)]*

2 + (2 + 2 + 2 + 2 + 2) = 12

Group - C

4. (a) Suppose that a data warehouse for Mega-Corporation consists of the following four dimensions: employee, project, quarter, and manager, and two measures count and avg_performance. When at the lowest conceptual level (e.g., for a given employee, project, quarter, and manager combination), the avg_performance measure stores the actual performance rating of the employee. At higher conceptual levels, avg_performance stores the average performance rating for the given combination.
(i) Draw a snowflake schema diagram for the data warehouse. *[(CO4)(Understand/IOCQ)]*
(ii) Starting with the base cuboid [employee, project, quarter, manager], what specific OLAP operations (e.g., roll-up from quarter to year) should one perform in order to list the average performance rating of employees on each project managed by a specific manager in Mega-Corporation? *[(CO4)(Understand/IOCQ)]*

(b) The sales for New York, Los Angeles, Chicago, and Houston are shown for the time dimension (organized in quarters) and the product dimension (classified according to the types of products sold) are displayed in the 2D table below. The fact or measure is displayed in revenue_sold (in thousands). Create slices of locations to view the data. *[(CO4)(Understand/IOCQ)]*

Quarter	Product Type	New York	Los Angeles	Chicago	Houston
Q1	Electronics	250	180	200	150
Q2	Electronics	300	220	230	180
Q3	Electronics	280	200	210	160
Q4	Electronics	320	240	250	190
Q1	Apparel	150	100	120	90
Q2	Apparel	180	120	140	110
Q3	Apparel	160	110	130	100
Q4	Apparel	200	140	160	120

(c) What is a star schema in data warehousing? *[(CO4)(Analyse/LOCQ)]*

(4 + 2) + 4 + 2 = 12

5. Suppose that a data warehouse for Big University consists of four dimensions: *student, course, semester,* and *instructor*, and two measures *count* and *avg_grade*.
   (a) Draw a snowflake schema diagram for the data warehouse. *[CSEN3237.4(Understand/IOCQ)]*
   (b) 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 Data Science courses for each Big University student? *[CSEN3237.4(Understand/IOCQ)]*
   (c) If each dimension has five levels (including all), such as "*student < course < semester < instructor < all*", how many cuboids will this cube contain (including the base and apex cuboids)? *[CSEN3237.4(Understand/IOCQ)]*

6 + 3 + 3 = 12

Group - D

6. (a) Outline the steps involved in creating a sample data warehouse, from conceptualization to implementation. Discuss key considerations for designing a data warehouse architecture. *[(CO3)(Apply/HOCQ)]*
   (b) Introduce materialized views and provide guidelines for designing their schema. Explain the concept of query rewrite and how it is related to materialized views. *[(CO4)(Remember/LOCQ)]*
   (c) Outline the role of integrity constraints in maintaining data consistency. Discuss the benefits and trade-offs associated with table compression in a data warehouse environment. *[(CO3)(Apply/HOCQ)]*

(2 + 2) + (2 + 2) + (2 + 2) = 12

7. (a) What is data partitioning? Give two reasons why data partitioning is helpful in a data warehousing environment. *[CSEN3237.3(Appreciate/IOCQ)]*
   (b) Determine the Data Partitioning Scheme by establishing Clustering Options, preparing an Indexing Strategy and assigning Storage Structures. *[CSEN3237.3(Appreciate/IOCQ)]*

(2 + 4) + 6 = 12

Group - E

8. (a) Outline the basic tasks involved in data cleansing during the ETL process. Discuss the significance of data cleansing and provide examples of common issues that may arise in raw data, emphasizing the importance of addressing these issues in the cleansing phase. *[(CO3)(Analyse/HOCQ)]*
   (b) Explain the importance of transportation in data warehouses. Discuss the procedures used for loading data into both Dimension and Fact tables, emphasizing key considerations and best practices. *[(CO6)(Remember/LOCQ)]*
   (c) Compare and contrast ROLAP (Relational OLAP), MOLAP (Multidimensional OLAP), and HOLAP (Hybrid OLAP) in terms of architecture and performance. Highlight the strengths and weaknesses of each architecture. *[(CO1)(Apply/IOCQ)]*
   (d) Briefly explain the importance of source identification in the extraction process and provide one example of a challenge that may arise during this phase. *[(CO4)(Apply/IOCQ)]*

3 + 3 + 4 + 2= 12

9. You are the staging area expert on the data warehouse project team for a large toy manufacturer. Discuss the four modes of applying data to the data warehouse. Select the modes you want to use for your data warehouse and explain the reasons for your selection. *[CSEN3237.6 (Develop/HOCQ)]*

12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	28.12	41.67	30.21

Course Outcome (CO):

After the completion of the course students will be able to

CSEN3237.1. Learn and understand various terminologies used for analytic data processing and understand the difference between OLTP and OLAP, RDBMS and data warehouse.
CSEN3237.2. Understand the functionality of the various data warehousing component.
CSEN3237.3. Appreciate the strengths and limitations of various data warehousing models.
CSEN3237.4. Understand the dimensional modeling technique and apply it for creating conceptual model.
CSEN3237.5. Develop skill in Special SQL syntax for physical implementation of Data warehouse.
CSEN3237.6. Develop skill through case study of some documented data warehouse model.

**LOCQ: Lower Order Cognitive Question; IOCQ: Intermediate Order Cognitive Question; HOCQ: Higher Order Cognitive Question.*