National University of Computer and Emerging Sciences 0042

Lahore Campus

Data Warehousing and Business Intelligence (DS3003)

Final Exam

Total Time (Hrs.): 3
Total Marks: 60
Total Questions: 5

Date: December 28th 2024

Course Instructor

M. Ishaq Raza

22L-7558 Roll No

BDS-5A Section

Student Signature

Do not write below this line.

Note: Please ensure that you attempt all questions and their respective parts in the given order. You may use a calculator.

CLO # 2: Demonstrate an understanding of the fundamental concepts of the Star and the Snowflake Schema; learn how to design the schema of a DW based on these two models.

Q. No 1: [12+4= 16]

a. As the data design specialist for the data warehouse project team of an eCommerce system, **design a star schema** that includes a base fact table with at least four dimensions and three aggregate fact tables (a 3-way, 2-way, and 1-way aggregation). Provide the possible attributes for each dimension and fact table. Show the primary keys, foreign keys and all the relationships between the dimension and fact tables. Additionally, design a dimension table that supports the preservation of historical changes. Note: Draw a single diagram that includes the base fact table and the aggregate fact tables.

b. Take appropriate cardinality (i.e., number of rows) of each of the above dimension and their levels. Estimate the potential size (in number of rows) of the above base fact table as well as aggregate fact tables.

CLO # 1: Demonstrate an appreciation of the role that DW and BI play in enhancing the decision-making process.

Q. No 2: Give the appropriate answers to the following questions: [15]

a. Briefly discuss the three major types of architectures for building a data warehouse.
b. Give at least three reasons why you think ETL functions are most challenging in a data warehouse environment.
c. Drill-down, drill-across, and drill-through are analytical operations that can be performed on an OLAP cube. Provide a brief description of each.
d. Discuss at least three advantages of using materialized views.
e. Can we use a combination of vertical and horizontal partition techniques together for optimal performance? Illustrate your answer with an example.

CamScanner

CLO # 1: Demonstrate an appreciation of the role that DW and BI play in enhancing the decision-making process.

Q. No 3: Suppose you have the following market basket data. [5]

TID	Items-Bought
1	{A, C, F}
2	{B, C, D}
3	{A, B, C, D}
4	{B, D}
5	{E}

Find all frequent itemsets using Apriori algorithm with min_sup=2, i.e., any itemset occurring in less than 2 transactions is infrequent. Also list all the strong association rules with min_sup=2 and min_conf=100%.

Consider the following description for the next two Questions:
Consider the following tables and statistics which are part of a vehicle sales system:
Vehicle *(VehicleID, Make, Model, Color, ...);*
Sales *(SalesID, VehicleID, SalesPersonID, SalesDate, Price, ...);*

Assume Vehicle and Sales tables containing 200,000 *and 150,000* rows respectively. Each table row and each index entry take 100 *bytes* and 20 *bytes of* space respectively. Data block size is 16*KB* and available memory size is *50 blocks*. Suppose selectivity of Model '2024'= 2%, Model '2023'= 4%, Color 'White'=8, and Color 'Black'=6%.

CLO # 1: Demonstrate an appreciation of the role that DW and BI play in enhancing the decision-making process.

Q. No 4: Calculate the total I/O cost to execute this Query using the following indexed access paths. Show all steps clearly. [12]

Query: *SELECT * FROM vehicle*
WHERE model IN (2024, 2023) AND color IN ('White', 'Black');

a. Single Index access *(Assume single indexes exist on model and color columns separately)*
b. Dynamic Bitmap Index access *(Assume single indexes exist on model and color columns separately)*
c. Composite Index Access (Assume a composite index exist on model and color columns, with index entry size=20 bytes)

CLO # 1: Demonstrate an appreciation of the role that DW and BI play in enhancing the decision-making process.

Q. No 5: Calculate the total I/O cost to execute this Query using the following joining techniques. Show all steps clearly. Assume there is no index exist on any table. [12]

Query: *SELECT * FROM vehicle JOIN sales ON vehicle.vehicleID = sales.vehicleID*
WHERE model IN (2024, 2023) AND color IN ('White', 'Black');

a. Hash Join
b. Sort Merge Join
c. Nested Loop Join *(Identify the most efficient variant of NLJ in this scenario, then compute the I/O cost of that variant only.)*

CamScanner