National University of Computer and Emerging Sciences

Lahore Campus

Data Warehousing and Business Intelligence (DS3003)

Sessional-I Exam

Total Time (Hrs.): 1

Total Marks: 25

Total Questions: 3

Date: September 23rd 2024

Course Instructor(s)

M. Ishaq Raza

_____ _____ _____

Roll No Section Student Signature

Solution

Do not write below this line.

Note: Please ensure that you attempt all questions and their respective parts in the given order.

Consider the following case study for the next two questions:

Bill Date Dim: Bill Date, Bill Day Desc, Bill Month ID, Bill Month Desc, Bill Year ID, Bill Year Desc

Customer Dim: Customer Code, Customer Desc, City ID, City Desc, Country ID, Country Desc

Sales Rep Dim: Sales Rep No, Sales Rep Desc, Channel ID, Channel Desc

Rate Plan Dim: Rate Plan Id, Rate Plan Desc, Rate Plan Type Code, Rate Plan Type Desc

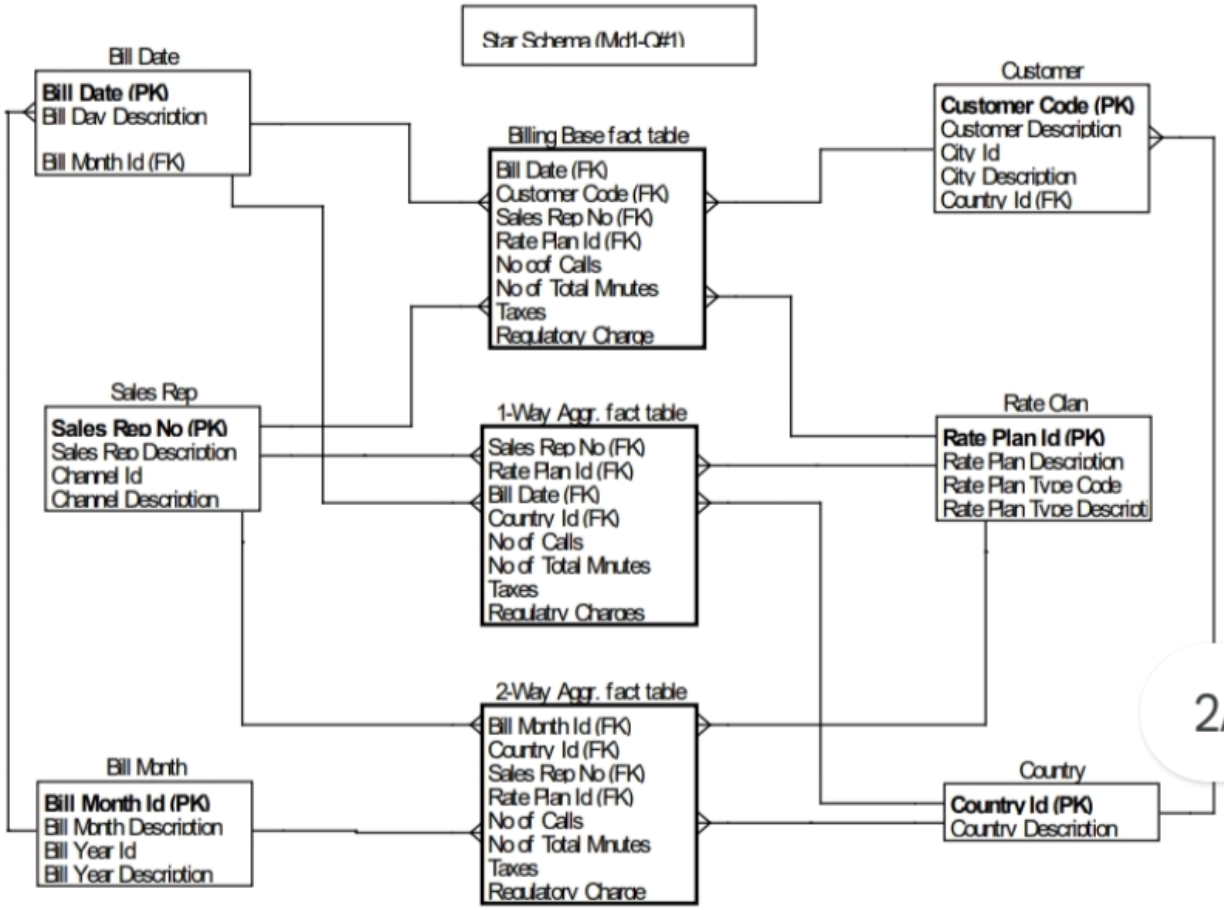
Billing Fact: Bill Date, Customer Code, Sales Rep No, Rate Plan Id, No of Calls, No of Total Minutes, Taxes, Regulatory Charge

Assume: 10,000 customers, 200 cities, 5 countries, 40 sales rep, 4 channels, 30 rate plans, 3 rate plan types, and 3 years billing history.

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: Draw the appropriate star schema that includes a base fact table, a one-way aggregate fact table (along Customer Country), and a two-way aggregate fact table (along Bill Month and Customer Country). Show the primary keys, foreign keys and all the relationships between the dimensions and fact tables. Note: Draw only one diagram that includes base fact table as well as aggregate fact tables. [10]

Ans:



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 2: Estimate the size (in number of rows) of the above customer dimension table, sales rep dimension table, billing base fact table, and both the aggregate fact tables. [5]

Ans:

Customer Dimension: 10000+5= 10005 rows

Sales Rep Dimension: 40 rows

Base Fact Table: (3x365) x 10000 x 40 x 30 = 13,140,000,000 rows

Aggregate Fact Table1 (along country:

3x365 (day) x 5 (country) x 40 (sales rep) x 30 (rate plan) = 6,570,000 rows

Aggregate Fact Table2 (along month & country):

3x12 (month) x 5 (country) x 40 (sales rep) x 30 (rate plan) = 216,000 rows

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 3: Briefly answer the following questions. [10]

a. Pick any one architecture for building a data warehouse and list the advantages and disadvantages of that architecture.

b. What are the different types of OLAP? Which type of OLAP can handle large amounts of data? Justify your answer.

c. How does a snowflake schema differ from a star schema? Name two advantages of the snowflake schema.

d. Differentiate between pre-join denormalization and column-replication denormalization techniques. Explain with an example.

e. When would you use partitioned cubes in multidimensional online analytical processing (MOLAP)?

Ans: See Textbook/Notes.

Q1)

Bill Date Dim:	Customer Dim:	Sales Rep Dim:
Bill Date	Customer Code	Sales Rep No
Bill Day Desc	Customer Desc	Sales Rep Desc
Bill Month ID	City ID	Channel ID
Bill Month Desc	City Desc	Channel Desc
Bill Year ID	Country ID	
Bill Year Desc	Country Desc	

Rate Plan Dim:
Rateplan ID
Rate plan Desc
Rate Plan Type Code
Rate Plan Type Desc

Billing ~~Dat~~ Fact:
Bill Date
Customer Code
Sales Rep No
Rate Plan ID
No of Calls
No. of total Minutes
Taxes
Regulatory charge

→ 10,000 customers
200 cities
5 countries
40 sales rep
4 channels
30 rate plans
3 rate plan types
3 years billing history

1 year 12

Q2)

Customer Dim: 10,000 rows

Sales Rep Dim: 40 rows

Billing Base Fact: $3 \times 365 \times 10{,}000 \times 40 \times 30$
$= 1.314 \times 10^{10}$ rows

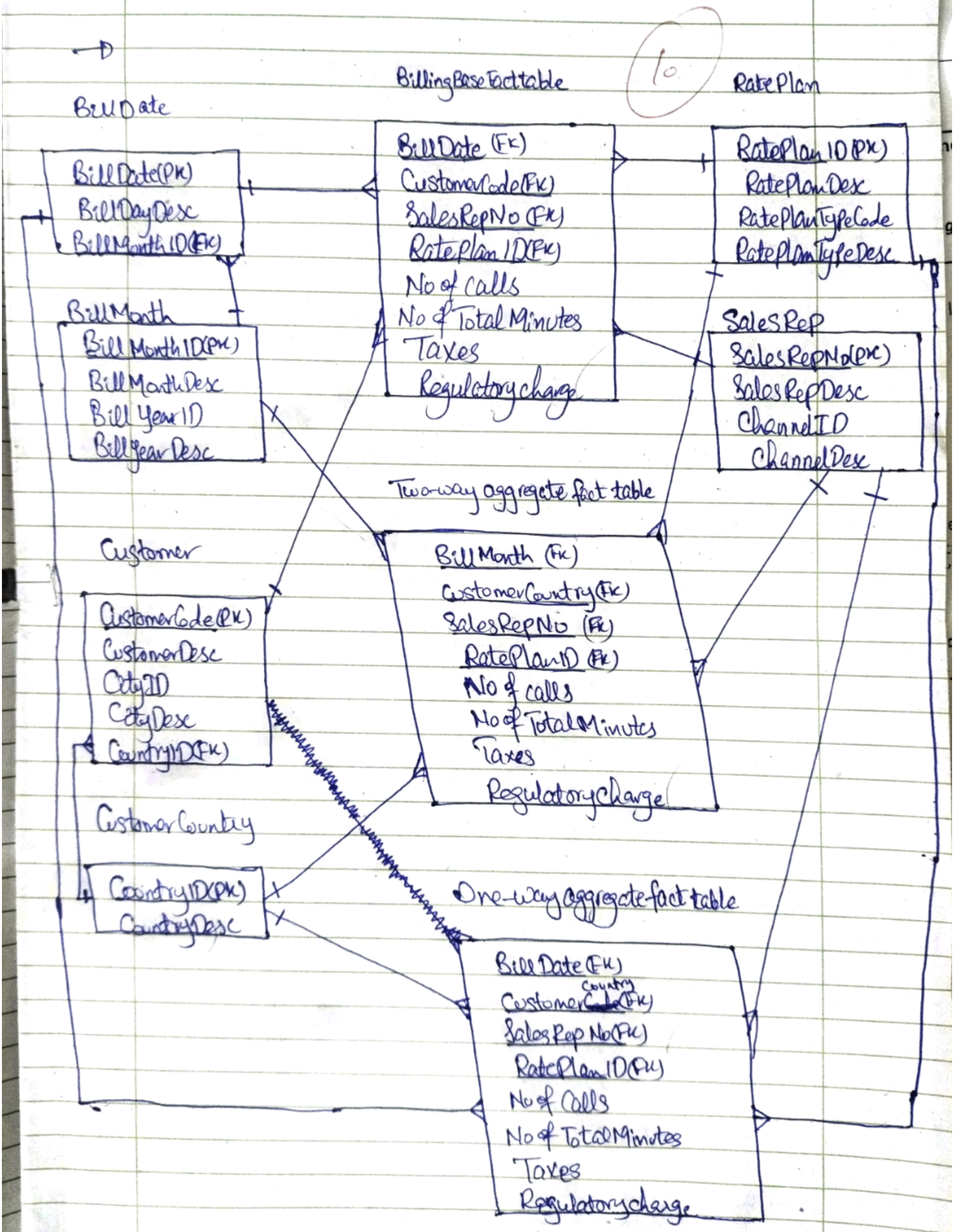
Two-way aggregate ~~fact table~~: $5 \times 40 \times 30 \times 3 \times 12$
$= 216000$ rows

One-way aggregate fact table: $3 \times 365 \times$ ~~10,000~~ $5 \times 40 \times$ ~~30~~ 30
$= 6570000$ rows

Q / Part No.

Star Schema

Rough Work



Q / Part No.

Rough Work

Q3

(a) Dependent data marts directly create data from centralized datawarehouse. They ensure data consistency across enterprise. Whereas, Independent data marts are standalone systems that create data from various resources without relying on a central data warehouse and hence result in a data inconsistency issues. There are various architectures like hub and spoke, federated, data mart bus, Centralized datawarehouse and independent data marts.

(c) Snowflake schema is normalized version of star schema where dimension tables are split into multiple related tables whereas, star schema is a simpler structure where central base fact table is connected to denormalized dimension tables.
Advantage: ① Minimal data redundancy ② Storage is reduced.

(d) Pre-join denormalization is a technique involving joining tables in advance and storing results in a single table to improve query performance. Whereas, column replication is a technique of replicating columns across multiple tables to avoid joins and simplified queries for better query performance. Since you do not need to join tables at query time, so it improves overall performance.

(e) Partitioned cubes are large OLAP cubes that are divided into smaller, manageable (portions) for performance optimization. They are useful for larger datasets allowing faster query processing and improved load performance.

(b) There are 4 different types of OLAP:
MOLAP
ROLAP
DOLAP
HOLAP
Since ROLAP stores data in relational databases and have slower query performance as it generates query dynamically but has better scalability on larger datasets.

Data Warehousing and Business Intelligence (DS3003)

Sessional-II Exam

Total Time (Hrs.): 1
Total Marks: 25
Total Questions: 3

Date: November 4th 2024

Course Instructor

M. Ishaq Raza

Roll No _____ 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 # 1: Demonstrate an appreciation of the role that DW and BI play in enhancing the decision-making process.

Q. No 1: Give the appropriate answers to the following questions: [3+2]

a. When it is appropriate to use full data refresh loading strategy? Also suggest at least two practical steps that expedite the full data refresh loading process.
b. Briefly explain the two most efficient immediate data extraction techniques.

Consider the following description for the nest Questions:

Consider the following tables and statistics which are part of a Library system:
Book *(BookID, Title, Publisher, PublishYear, Author, ...);*
BookLoan *(BookID, BranchID, CardNo, DateOut, DueDate, ...);*

Assume Book and BookLoan tables containing 100,000 *and 2,000,000* rows respectively. Each table row and each index entry take 200 *bytes* and 15 *bytes of* space respectively. Data block size is 8*KB* and available memory size is *100 blocks*. Suppose selectivity of publisher 'Wiley'= 10%, publishYear '2023'= 8%, author 'Inmon'= 4%, and author 'Kimball'=2%.

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

Q. No 2: Calculate the total I/O cost for the Query using the following indexed access paths. Show all steps clearly. [10]

Query: *SELECT * FROM book*
WHERE publisher= 'Wiley' AND publishYear= '2023' AND author IN ('Inmon', 'Kimball');

a. Clustered Index access *(Assume clustered index exit on Author column)*
b. Static Bitmap Index access *(Assume static bitmap indexes exist on Publisher, PublishYear, and Author columns separately)*

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

Q. No 3: Calculate the total I/O cost to execute this Query using the following joining techniques. Show all steps clearly. Assume static bitmap indexes exist on Publisher, PublishYear, and Author columns of Book table and there is no index on BookLoan table. [3+3+4= 10]

Query: *SELECT * FROM book JOIN BookLoan ON Book.BookID = BookLoan.BookID*
WHERE publisher= 'Wiley' AND publishYear= '2023' AND author IN ('Inmon', 'Kimball');

a. Sort Merge Join
b. Hash 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.)*

Student's Name Ameer Abdullah Signature

Roll No 21L-5694 Section BDS-5A Date 4th Nov, 2024

Q1)

a) We use full data refresh loading strategy when our large data changes occur - like the ratio between existing and new data size exceeds 10%. The steps to expedite the full data refresh is (i) Batch processing (ii) parallel processing (iii) Drop and rebuild indexes. 3

b) (i) Transaction log: They are already in relational database management and does not require extensive customization. It is one of the most cost effective technique. 3

(ii) Database Triggers: They are built in database and automatically track changes without requiring any additional tools or processes. They are easy to setup and manage. 2

Q2)

$r_B = 100000$
$r_{BL} = 2000000$
$R = 200$ bytes
$R_i = 15$ bytes
$B = 8KB \rightarrow 8 \times 1024$
$K = 100$ blocks
Publisher Wiley = 10%
PublishYear 2023 = 8%
Author Inmon = 4%
Author Kimball = 2%

$bfr = \left\lfloor \frac{8 \times 1024}{200} \right\rfloor = 40$ blocks

$bfr_i = \left\lfloor \frac{8 \times 1024}{15} \right\rfloor = 546$ index entries

$b_B = \left\lceil \frac{100000}{40} \right\rceil = 2500$ blocks

$b_{BL} = \left\lceil \frac{2000000}{40} \right\rceil = 50000$ blocks

1:20

Q / Part No.

Rough Work

$b_{Bi} = \left\lceil \frac{100000}{546} \right\rceil = 184$ index blocks

$b_{BLi} = \left\lceil \frac{2000000}{546} \right\rceil = 3664$ index blocks

(1c)

Qualifying rows = 10% × 8% × 6% × 100000 ⇒ 48

Qualifying blocks: $\left\lceil \frac{48}{40} \right\rceil = 2$ blocks

(Q2)

(a) Clustered Index access:

6% × 100000 = 6000

Base table access cost: $\left\lceil \frac{6000}{40} \right\rceil = 150$

Index table access cost: $\left\lceil \frac{6000}{546} \right\rceil = 11$

Total cost (Dense Index): 150 + 11 = 161 I/O

→ Sparse Index → Index table access cost: 1 + 1 = 2

→ 150 + 2 = 152 I/O

(b) $\left\lceil \frac{100000}{8 \times 15 \times 546} \right\rceil = 2$ (Size for one bitmap)

(2 + 2 + 2 + 2) = 8

Since QR < b_B

48 < 2500

So, Total Cost: 8 + 48 = 56 I/O

Q3

QB < K] so we do not ~~merge~~ sort Book table

(a) SMJ: & 2 < 100]

FTS + (Book Sort) + (Book Loan Sort) + (Merge Cost)

2500 + 2 + $\left(50000 \times \log_2\left(\frac{50000}{100}\right)\right)$ + (2 + 50000)

~~2500~~ 56 + 2 + 282193 + 50002

2.5

Total I/O cost : 334697 I/O

we took FTS = 56 due to static bitmap index on outer table.

(b) Hash Join: NO Partitioning cost for both table as smaller table can fit into the memory.

FTS + Hashing = Total I/O cost

2500 + (2 + 50000) = 52502 I/O

2.5

(c) Nested Loop Join:

Best variant:

Clustered NLJ: FTS + QR × (Index access cost + base table access cost)

: 2500 + 48 × (1 + 1) = 2596 I/O

no index exist on inner table!

2nd Best variant:

Indexed NLJ: 2500 + 48 × (1 + 20) = 3508 I/O

→ Since Basic and Block NLJ will be highly costly so those are not good option.

→ Also as we can see that Indexed NLJ cost is significantly greater so we consider clustered NLJ as the best.

* As bitmap indexes exist ~~on~~ so 56 + 48(1+1) = 152 I/O *

(c) 56 + (2 × 50000) correct ✓

National University of Computer and Emerging Sciences 0002

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

Roll No _____ 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.

Product ID
Sale ID
Customer ID
Billing Date
Total Sale
Taxes
Products sold overtime
No. of customers

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.

$\frac{2\% \times 8\% \times 200000}{819} = 1+320$

$\frac{2\% \times 6\% \times 200000}{819} = 1+240$

$\frac{4\% \times 8\% \times 200000}{819} = 1+640$

$\frac{4\% \times 6\% \times 200000}{819} = 1+480$

$\rightarrow 1684$

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%.

BC→D A→C D→B
CD→B B→D

BC→D B→D
CD→B D→B
A→C

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)* 1242
b. Dynamic Bitmap Index access *(Assume single indexes exist on model and color columns separately)* 1277
c. Composite Index Access (Assume a composite index exist on model and color columns, with index entry size=20 bytes) 1684

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 2159 or 2156
b. Sort Merge Join 6042
c. Nested Loop Join *(Identify the most efficient variant of NLJ in this scenario, then compute the I/O cost of that variant only.* 113581o 11358