Re-Assessor

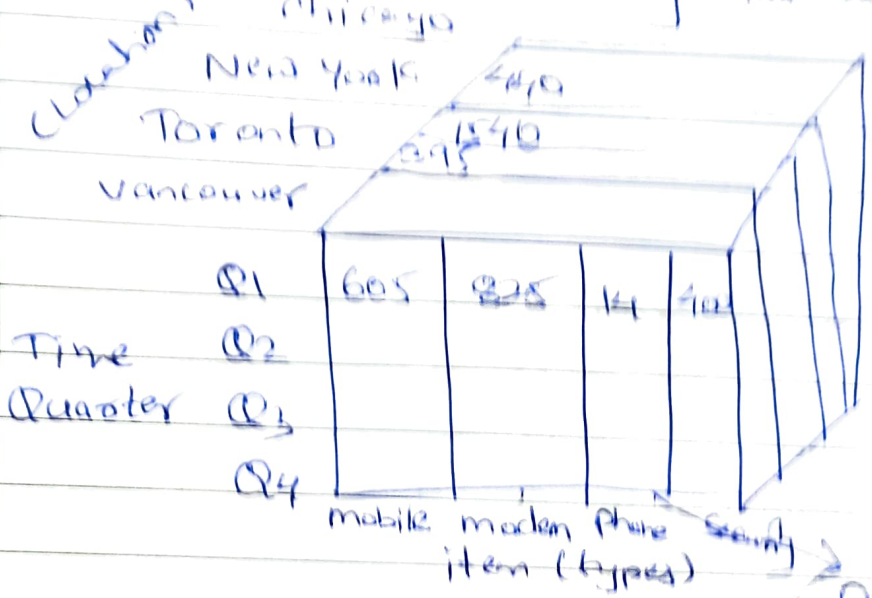
ce for arks | Question No.

START WRITING HERE

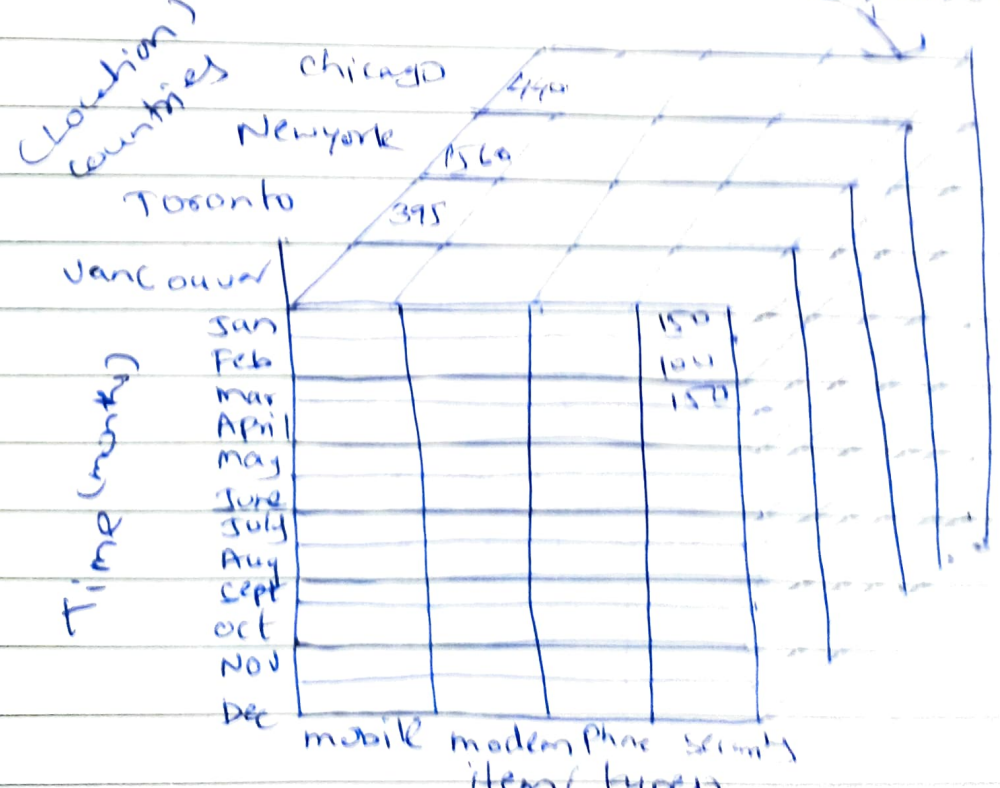
Q1 Attempt any five (5 marks)

(a) Drill down operation

Drill down operation is reverse of the roll-up operation. i.e by stepping down a concept hierarchy or introducing




Drill down on time (from quarters to months)



VIT

(Autonomous College Affiliated to University of Mumbai)

Date: 21 Nov 2022

Semester: V

End Semester Examination (AY 2022-, Winter, 2022-23)

Branch: Comp Sc Engineering

Subject: Data Warehousing and Mining

Duration: 2½ Hrs

Marks: 60

N.B. - All Questions are Compulsory

| | | | |
|---|---|---|---|
| Q 1) | Attempt any Five (5 Marks Each) | | |
| a) | Explain Drill Down operation with example | CO1 | U |
| b) | Explain the features of a Data Warehouse | CO1 | U |
| c) | Explain the factless fact table | CO2 | U |
| d) | Explain the Dendrogram with respect to Clustering algorithms | CO4 | U |
| e) | Explain Support and Confidence with examples | CO4 | U |
| f) | Explain the Apriori principle | CO4 | U |
| g) | Explain Market Basket Analysis | CO4 | U |
| h) | Explain the applications of Web Mining | CO5 | U |
| Q 2) | Attempt any One (10 Marks Each) | | |
| a) | Explain typical OLAP operations with an example. | CO2 | U |
| b) | Consider a transactional dataset: | CO4 | A |

| TID | Items |
|---|---|
| T1 | I1, I2, I4 |
| T2 | I2, I3, I4 |
| T3 | I1, I2 |
| T4 | I2, I4 |
| T5 | I1, I2, I3 |

If minimum support = 30% and minimum confidence = 75%, find all possible association rules using the Apriori algorithm.

| | | | |
|---|---|---|---|
| Q 3) | Attempt any One (10 Marks Each) | | |
| a) | Draw the Star and Snowflake Schema for "Hotel Occupancy" considering dimensions like Time, Hotel, Room, etc. | CO2 | A |
| b) | Suppose that the data mining task is to cluster points [illegible] into 3 clusters. B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9) representing coordinates of location (say) into 3 clusters. Suppose initially A1, B1, and C1 are represented as the centers of each cluster respectively. Use Euclidean distance and apply the *k-means* algorithm to show: a. The three cluster centers after the first round of execution. | CO4 | A |
| Q 4) | Attempt any two. (5 Marks Each) | | |
| a) | Compare OLAP and OLTP. | CO1 | U |
| b) | Explain the ETL process of data warehousing | CO2 | A |
| c) | Explain Page Rank algorithm in detail. | CO5 | U |
| Q 5) | Attempt any one. (10 Marks Each) | | |
| a) | Apply agglomerative algorithm for Hierarchical Clustering of the below spatial data points using single link distance and sketch the dendrogram: | CO4 | A |

| Data Point | X-Coordinate | Y-coordinate |
|---|---|---|
| A | 1 | 1 |
| B | 2 | 2 |
| C | 6 | 6 |
| D | 4 | 4 |
| E | 3 | 4 |

| | | | |
|---|---|---|---|
| b) | Explain three tier data warehouse architecture with a suitable block diagram | CO1 | U |

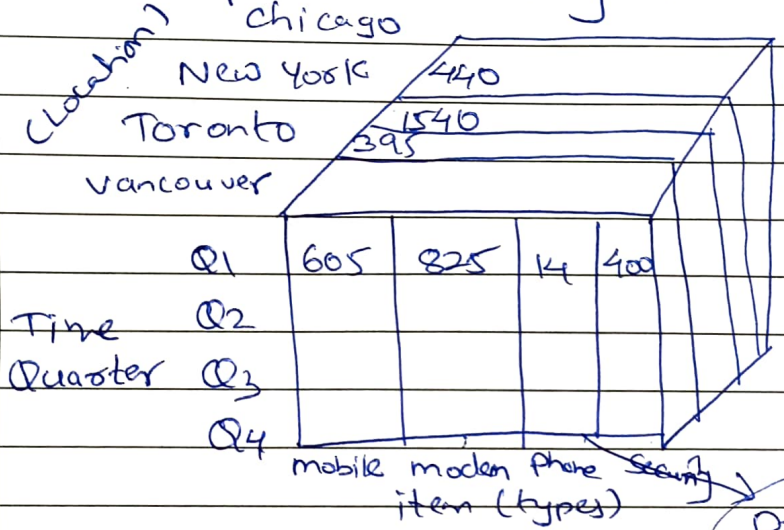
| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

| Space for Marks | Question No. | START WRITING HERE |
|---|---|---|

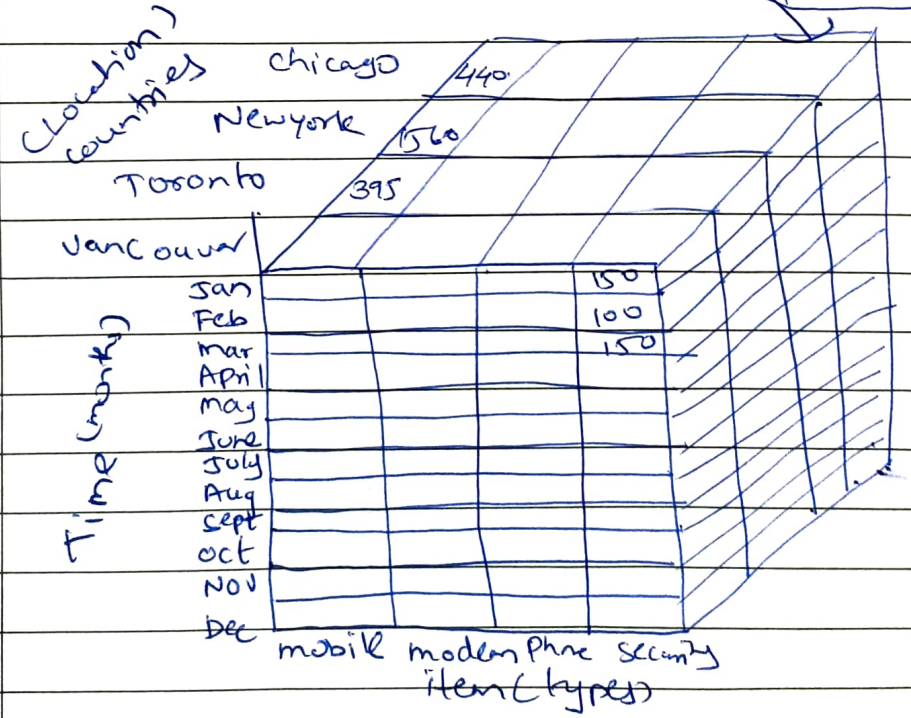
Q1(a) Attempt any five (2 marks)

(a) Drill down operation

Drill down operation is reverse of the roll-up operation i.e by stepping down a concept hierarchy for a dimension




Drill down on time (from quarters to months)



| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

| Space for Marks | Question No. | START WRITING HERE |
|---|---|---|

(b) Explain the features of Datawarehouse

→ 1) Deployed as a central database for the enterprise
2) Provide ETL (extract, transform, load) data processing capability
3) store meta data
4) Include access to reporting tools

(c) Explain factless fact table.

→ The fact table that do not have measures and has only the primary key of the dimensional table are called as factless fact table eg.
a record of student attendance in classes. In this case, the fact table would consist of 3 dimensions the student dimension the time dimension and the class dimension. This factless fact table would look like the following

Fact-attendance

| student_ID |
|---|
| class-ID |
| Time-ID |

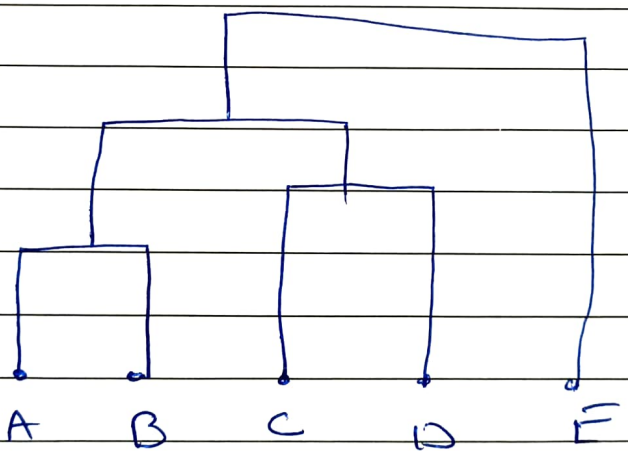
| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

| Space for Marks | Question No. | START WRITING HERE |
|---|---|---|

(d) Explain the dendrogram with respect to clustering algorithm

→ A tree of clusters that is formed after decomposing data objects into several levels of nested partitioning is called a dendrogram

A clustering of the data objects is obtained by cutting the dendrogram at the desired level. Then each connected component forms a cluster




(e) Explain support and confidence with example.

Support :- $\frac{\text{No of tuples that satisfy } x \text{ \& } y}{\text{Total No Tuples}}$

confidence = $\frac{\text{No of tuples that satisfy } x \text{ \& } y}{\text{Total No of tuples satisfy } x}$

| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

| Space for Marks | Question No. | START WRITING HERE |
|---|---|---|

Example TID Item bought

| TID | | Item bought |
|---|---|---|
| 10 | | milk bread cookies Juice |
| 10 | 20 | Milk, Juice |
| 20 | 30 | Milk, Eggs |
| 30 | 40 | Bread, cookies, coffee |
| 40 | | |

The support of Milk → Juice = 2/4 = 50%

The confidence of mik → Juice = 66.7%. 2/3

(f) Explain the Apriori principle

(1) The downward closure property of frequent patterns

→ Any subset of a frequent itemset must be frequent

→ If {milk, Bread, Juice} is frequent, so is {milk, Bread}

→ i.e every transaction having {milk, Bread, Juice} also contains {milk, bread}

(2) Apriori pruning principle: If there is any itemset which is infrequent, its superset should not be generated/tested.

| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

| Space for Marks | Question No. | START WRITING HERE |
|---|---|---|
| | | (1) Initially, scan DB to get frequent 1 itemset<br>(2) generate length (K+1) candidate itemsets from length K frequent itemsets<br>(3) Test the candidates against DB<br>(4) Terminate when no frequent or candidate set can be generated. |
| | (g) | Explain Market Basket Analysis<br>→ Market Basket Analysis goal is to understand consumer behavior by identifying relationship between the items that people buy. For example people who buy green tea are also likely to buy honey. So Market Basket Analysis would quantitiatively establish that there is a relationship between green tea and ~~bob~~ honey. |
| | (h) | Explain the application of web mining<br>→ (1) Web mining helps to improve the power of web search engine by classifying the web documents and identifying the web pages<br>(2) It is used for web searching |

| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

| Space for Marks | Question No. | START WRITING HERE |
|---|---|---|

eg: google yahoo etc and vertical searching eg. Fat leng etc.

3) It is used to predict user behavior

4) is useful of a particular website and e-service eg., landing page optimization.

Q Explain different OLAP operations with example.

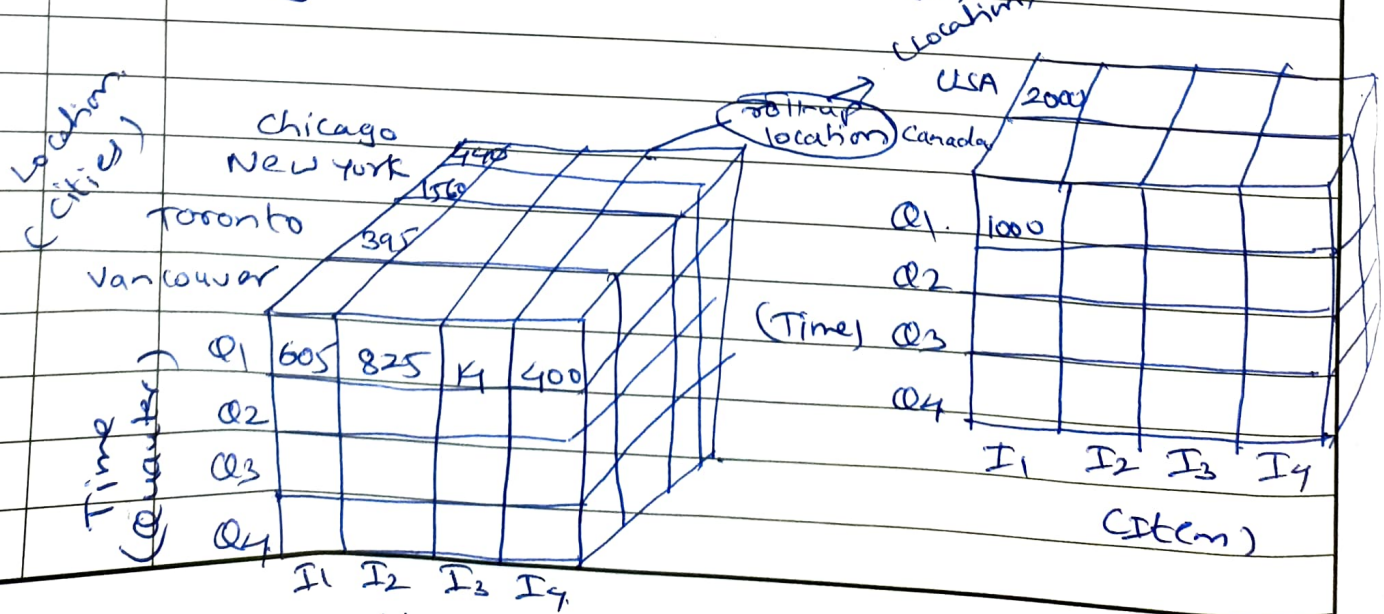
→ OLAP operation

(1) Roll-up

this operation performs aggregation on a data cube in any of the following way

(1) By climbing up a concept hierarchy for a dimension

(2) By dimension reduction



| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

| Space for Marks | Question No. | START WRITING HERE |
|---|---|---|

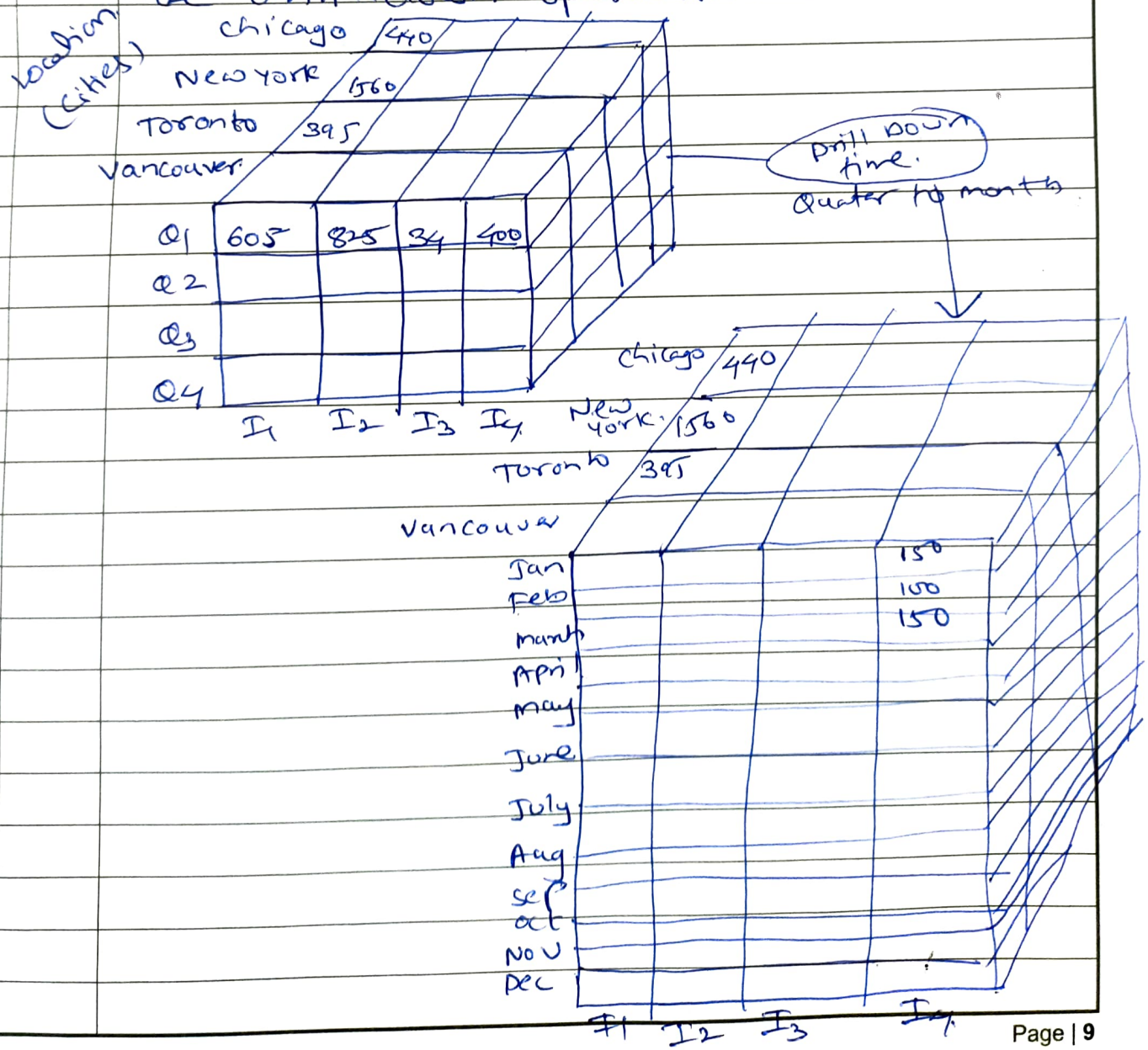
## Drill-down

drill-down operation is reverse of the roll-up. This operation is performed by either of the following way.

(1) By stepping down a concept hierarchy for a dimension

(2) By introducing new dimension

(3) Consider the following diagram showing the drill-down operation.

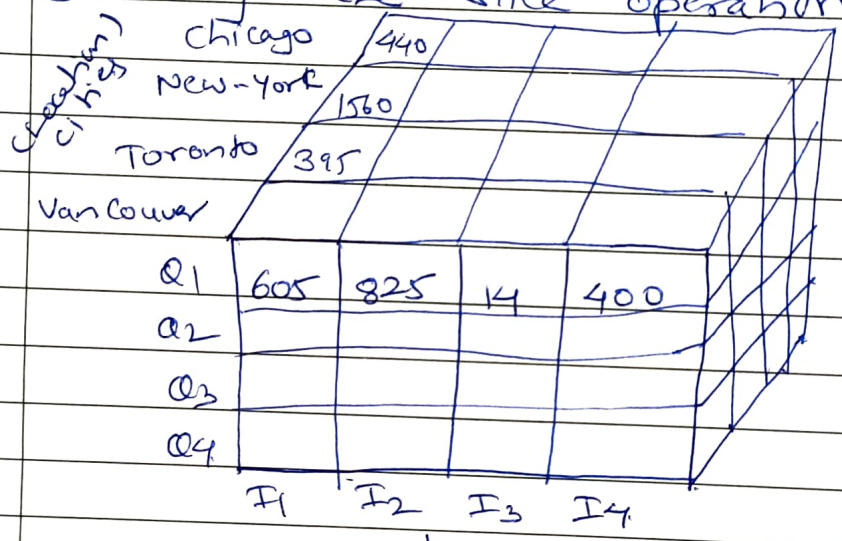


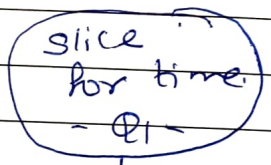
| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

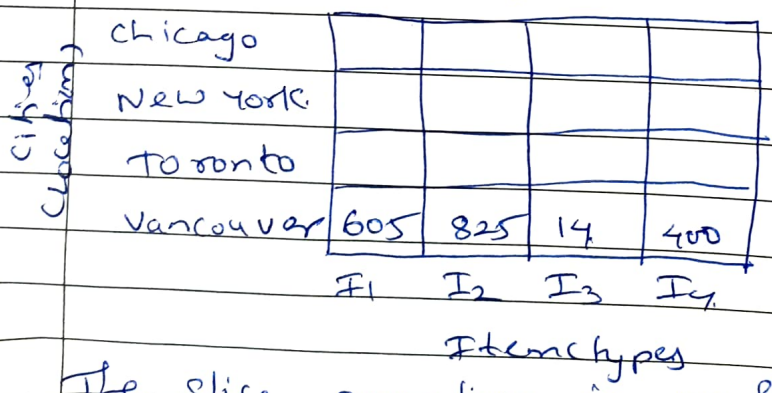
| Space for Marks | Question No. | START WRITING HERE |
|---|---|---|

## Slice

The slice operation performs selection of one dimension on a given cube and give us a new sub cube.
Consider the following diagram showing the slice operation.




The slice operation is performed for the dimension time using the criterion time = Q1'.

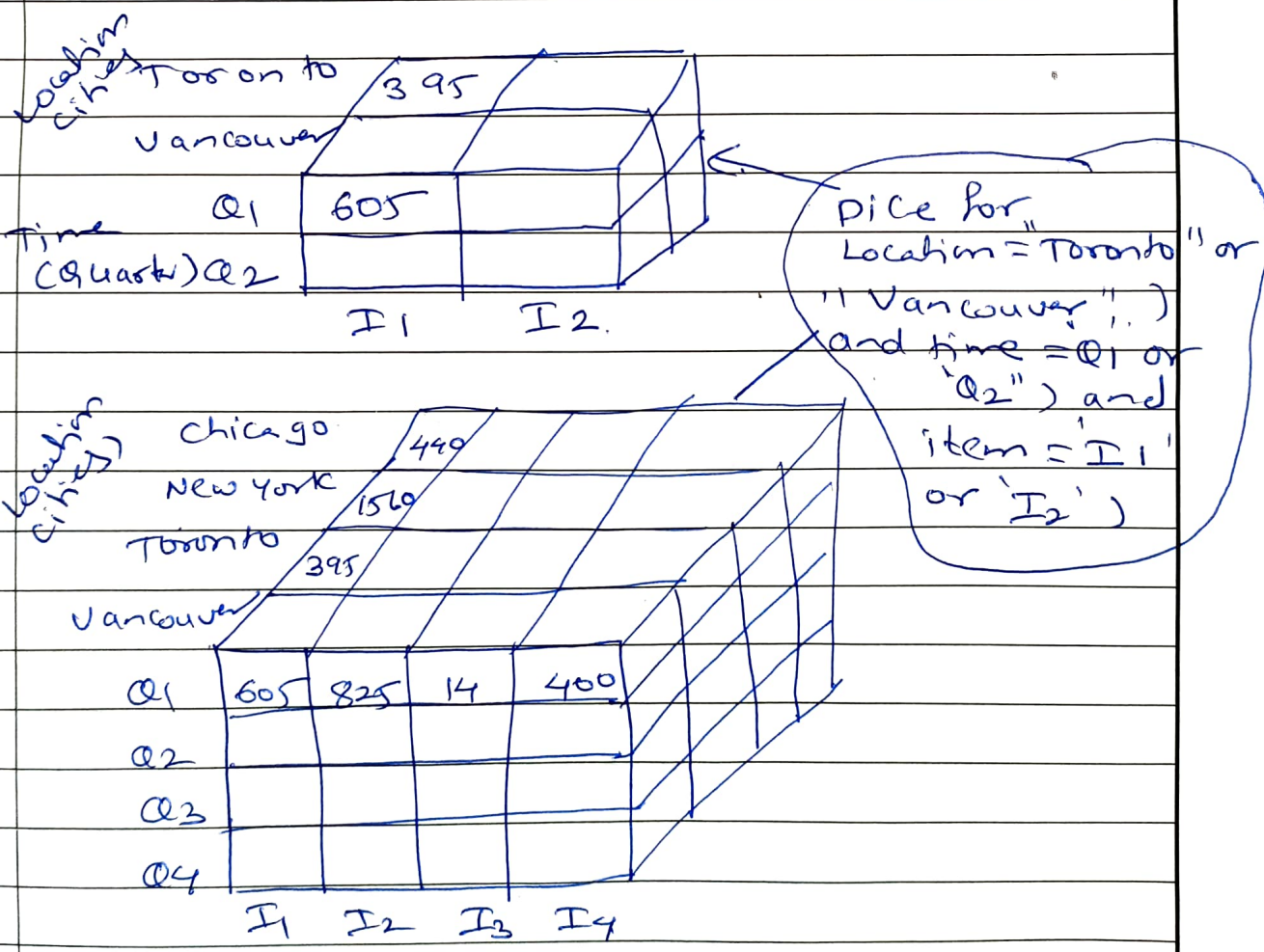
| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

| Space for Marks | Question No. | START WRITING HERE |
|---|---|---|

It will form a new sub cube by selecting one or more dimensions.

DICE

The dice operation performs selection of two or more dimension on a given cube and give us a new subcube Consider the following diagram showing the dice operation.



| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

| Space for Marks | Question No. | START WRITING HERE |
|---|---|---|

Q2

(b) Consider a transactional dataset

| TID | Items |
|---|---|
| T1 | I1, I2, I4 |
| T2 | I2, I3, I4 |
| T3 | I1, I2 |
| T4 | I2, I4 |
| T5 | I1, I2, I3 |

min support = 30% = $5 \times \frac{30}{100} = 1.5 \approx 2$

I^st Scan C1

| Itemset | Sup |
|---|---|
| {I1} | 3 |
| {I2} | 4 |
| {I3} | 2 |
| {I4} | 3 |

→ C2 = L1 ⋈ L1

| Itemset |
|---|
| {I1, I2} |
| {I1, I3} |
| {I1, I4} |
| {I2, I3} |
| {I2, I4} |
| {I3, I4} |

&

| Itemset | Sup |
|---|---|
| {I1, I2} | 3 |
| {I1, I3} | 1 |
| {I1, I4} | 1 |
| {I2, I3} | 2 |
| {I2, I4} | 3 |
| {I3, I4} | 1 |

L2

| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

| Space for Marks | Question No. |
|---|---|

START WRITING HERE

$C_3 \Rightarrow L_2 \bowtie L_2$

| Itemset |
|---|
| $\{I_1\ I_2\}$ |
| $\{I_2\ I_3\}$ |
| $\{I_2\ I_4\}$ |

$L_2$

$\rightarrow$

| Itemset | Sup |
|---|---|
| $I_1 I_2 I_3$ | 1 |
| $I_1 I_2 I_4$ | 1 |
| $I_2 I_3 I_4$ | 1 |

$C_3$

As $C_3$ has all the itemset whose minimum support is less than 2.

∴ The Frequent itemsets are

(1) $\{I_1\ I_2\}$ – 3

(2) $\{I_2\ I_3\}$ – 2

(3) $\{I_2\ I_4\}$ – 3

$i_1 \rightarrow i_2$ $3/3$

✗ $i_1 \rightarrow i_1$ $3/5$

✗ $i_2 \rightarrow i_3$ $2/5$

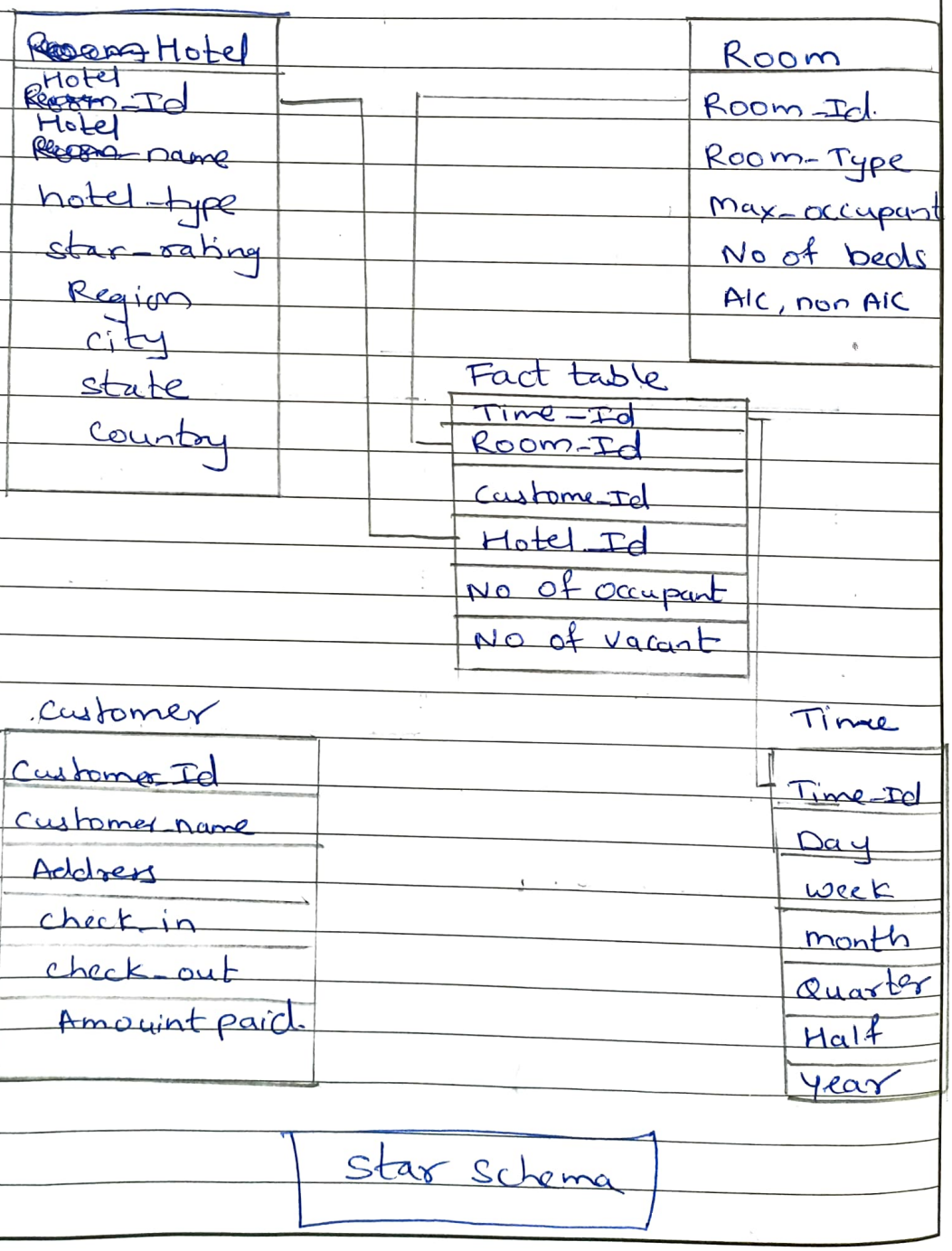
✓ $i_3 \rightarrow i_2$ $2/2$

✗ $i_2 \rightarrow i_4$ $3/5$

✓ $i_4 \rightarrow i_2$ $3/3$

| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

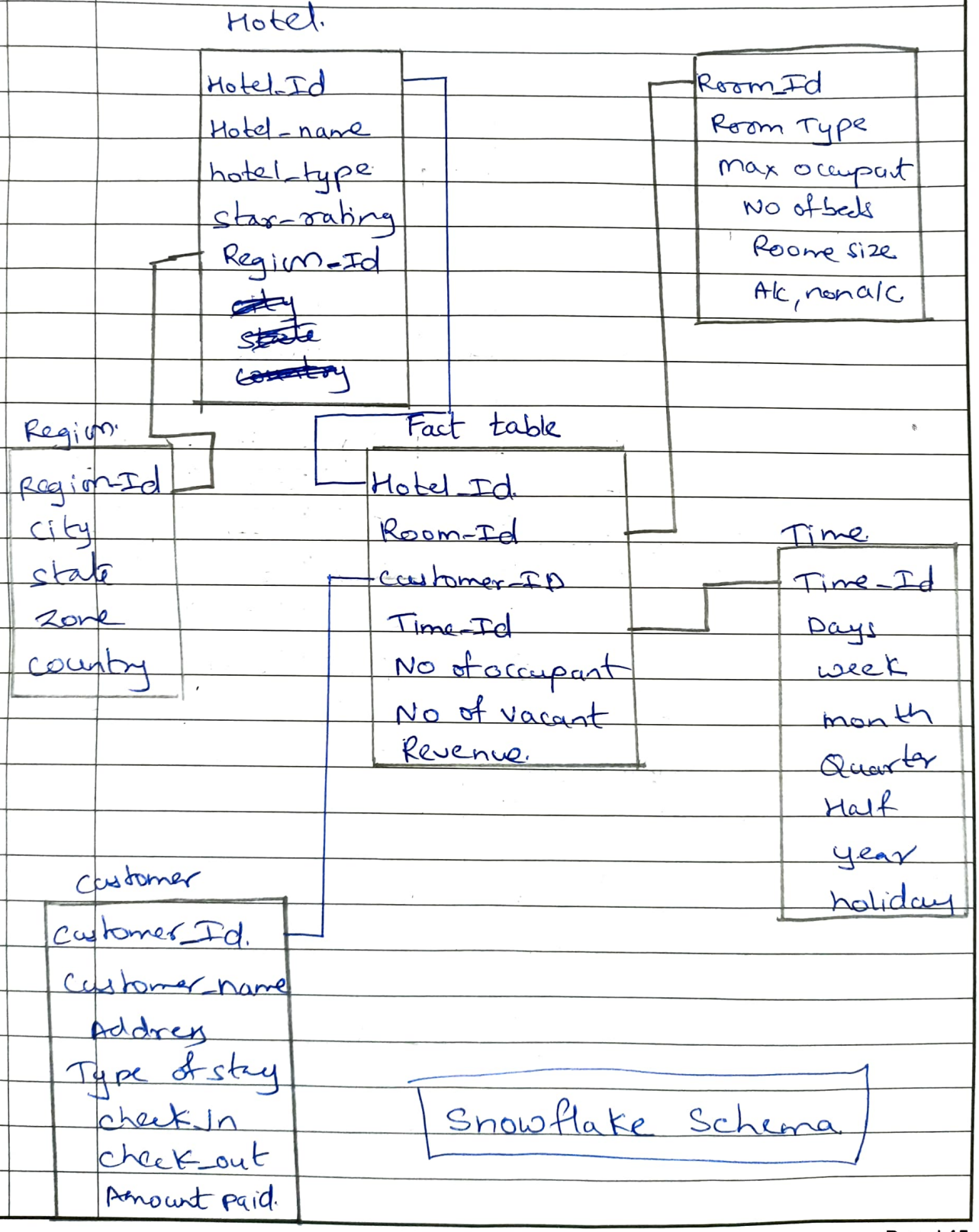
| Space for Marks | Question No. | START WRITING HERE |
|---|---|---|
| | Q3 a) | Draw the star and Snowflake schema for Hotel Occupancy considering dimension like time, hotel, Room etc. |



| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

| Space for Marks | Question No. | START WRITING HERE |
|---|---|---|

Q3(a)



| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

| Space for Marks | Question No. | START WRITING HERE |
|---|---|---|

(Q2(b))

Initial value of centroids

A1 B1 C1.

A1 (2,10) = D1

B1 (5,8) = D2

C1 (1,2) = D3

Now we calculate the distance between cluster centroid to each object by using euclidean distance

| | x | y | (A1) D1 | (B1) D2 | (C1) D3 | Cluster |
|---|---|---|---|---|---|---|
| A1 | 2 | 10 | 0 | 3.61 | 8.06 | 1 |
| A2 | 2 | 5 | 5 | 4.24 | 3.16 | 3 |
| A3 | 8 | 4 | 8.48 | 5 | 7.28 | 2 |
| B1 | 5 | 8 | 3.61 | 0 | 7.21 | 2 |
| B2 | 7 | 5 | 7.07 | 3.61 | 6.71 | 2 |
| B3 | 6 | 4 | 7.21 | 4.12 | 5.39 | 2 |
| C1 | 1 | 2 | 8.06 | 7.21 | 0 | 3 |
| C2 | 4 | 9 | 2.24 | 1.41 | 7.62 | 2 |

1) Final clusters are

Cluster1 = { A1, B1, C2 }

Cluster 2 = { A3 B2, B3 }

cluster 3 = { A2, C1 }

| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

| Space for Marks | Question No. | START WRITING HERE |
|---|---|---|

Q4(a) Compare OLAP & OLTP

| | OLTP | OLAP |
|---|---|---|
| (1) | OLTP stands for online transaction processing (OLTP) | OLAP stands for online analytical processing |
| (2) | OLTP works with processing of transaction | OLAP is more focused on analytical processing. |
| (3) | OLTP is Application oriented | OLAP is subject oriented. |
| (4) | The Usage is Repetive | The Usage is Ad-hoc |
| (5) | Users are in Thousands | Users are hundreds |
| (6) | Short Simple transaction | Complex Query |
| (7) | The data is current, up-to-date, detailed flat relational Isolated | Data is historical Summarized multidimensional Integrated & consolidated |
| (8) | Acess is Read/write Index/hash on primary key | Read Access only. |

| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

| Space for Marks | Question No. | START WRITING HERE |
|---|---|---|

Q4(b) Explain the ETL proces of data warehousing.

→ The ETL is abbreviation of Extract Transform and Load.

In this process an ETL tool extract the data from different RDBMS source systems then transforms the data like applying calculation, concentenation and load the data into the DW.

Step 1) Extraction :- In this step, data is extracted from the source system into the staging area. The data in the staging area are in form of flat files.

Staging area gives an apportunity to validate extracted data before it moves into the Datawarehouse. Source include legacy application like mainframes, POS like ATM spreed sheets, ERP. Hence we need to map the data before the data is extracted and loaded physically. This data map describes the relationship between sources and target data.

| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

| Space for Marks | Question No. | START WRITING HERE |
|---|---|---|

Three Data Extraction methods

(1) Full Extraction

(2) partial Extraction:- with update notification

(3) Partial Extraction:- without update notification

Irrespective of the method used, extraction should not affect performance and response time of the source system

Some validation are done during extraction

(1) make sure that no spam/ unwanted data is loaded.

(2) Remove all types of duplicate

(3) check all the keys are at place.

(2) Transformation

Data extracted from source server is raw and not usable in its original forms. Therefore it needs to be cleansed data that does not require any transformation is called an direct move

In transformation you can perform

(1) Customized operation

(2) Filtering

(3) Data standardization

| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

| Space for Marks | Question No. | START WRITING HERE |
|---|---|---|
| | 4 | Character set conversion and encoding handling |
| | 5) | Conversion of Unit |
| | (6) | Data threshold :- eg age cannot be greater than 2 digit. |
| | (7) | Cleaning |
| | (8) | split column or merge columns |
| | (9) | Transposing rows & Columns |
| | | step 3:- Loading:- It is last step of ETL proces. In a typical DW huge volume of data needs to be loaded in a relatively short periods. |
| | | Hence load proces should be optimized for performance |
| | | Types of loading |
| | (1) | Initial load :- populating all the DW tables |
| | 2) | Incremental load:- applying on going changes as when needed periodically |
| | 3) | Full Refresh:- Erasing the contents of one or more tables and reloading with fresh data. |

| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

| Space for Marks | Question No. | START WRITING HERE |
|---|---|---|
| | Q4c. | Explain Page Rank algorithm in detail. |
| | ⇒ | Page rank was proposed by Sergey Brin and Larry page students at Standford University and the founders of google. |
| | | Page rank is a numeric value that represents the importantance of a page present on the web. When one page links to another page, it is effectively cast in a vote for the other page. More votes implies |

| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

| Space for Marks | Question No. | START WRITING HERE |
|---|---|---|
| | Q4c. | Explain Page Rank algorithm in detail. |
| | → | Page rank was proposed by Sergey Brin and Larry page students at standford University and the foundes of google. |
| | | Page rank is a numeric value that represents the importantance of a page present on the web. When one page links to another page, it is effectively cast in a vote for the other page. more votes implies |

| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

| Space for Marks | Question No. | START WRITING HERE |
|---|---|---|
| | Q5(a) | Apply agglomerative algorithm for Hierarchical clustering of the above spatial data points using single link distance and sketch the dendrogram |

| Data Point | X-Coordinate | Y-Coordinate |
|---|---|---|
| A | 1 | 1 |
| B | 2 | 2 |
| C | 6 | 6 |
| D | 4 | 4 |
| E | 3 | 4 |

Using Euclidian distance Find distance matrix

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 0 | | | | |
| B | 1.4 | 0 | | | |
| C | 7.07 | 5.6 | 0 | | |
| D | 4.24 | 2.8 | 2.82 | 0 | |
| E | 3.60 | 2.23 | 3.60 | (1) | 0 |

$d(A,B) = \sqrt{(1-2)^2 + (2-1)^2} = \sqrt{2} = 1.4$

$d(A,C) = \sqrt{(6-1)^2 + (6-1)^2} = \sqrt{50} =$

$d(A,D) = \sqrt{(1-4)^2 + (1-4)^2} = \sqrt{18} =$

$d(A,E) = \sqrt{(1-3)^2 + (1-4)^2} = \sqrt{4+9} = \sqrt{13}$

$d(B,C) = \sqrt{(2-6)^2 + (2-6)^2} = \sqrt{32} = 5.$

$d(B,D) = \sqrt{(2-4)^2 + (2-4)^2} = \sqrt{8} =$

4 + 4

| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

Space for Marks | Question No.

START WRITING HERE

$d(B,E) = \sqrt{(2-3)^2 + (2-4)^2} \quad \sqrt{5}$

$d(C,D) = \sqrt{(6-4)^2 + (6-4)^2} = \sqrt{8}$

$d(C,E) = \sqrt{(6-3)^2 + (6-4)^2} = \sqrt{13}$

$d(D,E) = \sqrt{(4-3)^2 + (4-4)^2} = \sqrt{1} = 1$

d(D, E) is merged.

| | A | B | C | {D, E} |
|---|---|---|---|---|
| A | 0 | | | |
| B | 1.414 | 0 | | |
| C | 7.071 | 5.656 | 0 | |
| {D, E} | 3.605 | 2.236 | 2.828 | 0 |

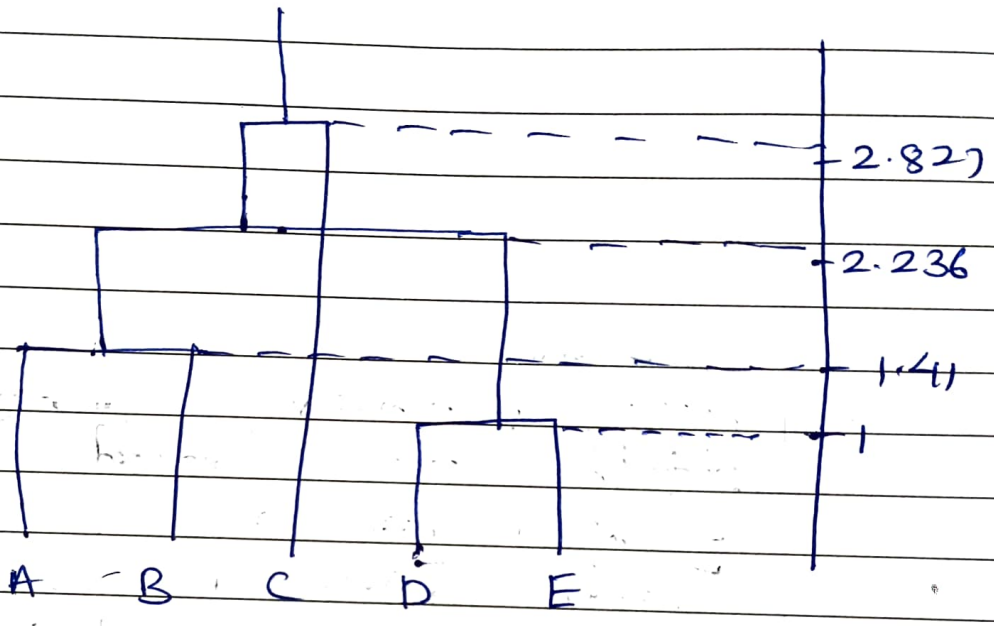
Since single link distance we will find minimum distance

| | {A, B} | C | {D, E} |
|---|---|---|---|
| {A, B} | 0 | | |
| C | 5.656 | 0 | |
| {D, E} | 2.236 | 2.828 | 0 |

| | {A, B, D, E} | C |
|---|---|---|
| {A, B, D, E} | 0 | |
| C | 2.828 | 0 |

| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

| Space for Marks | Question No. | START WRITING HERE |
|---|---|---|

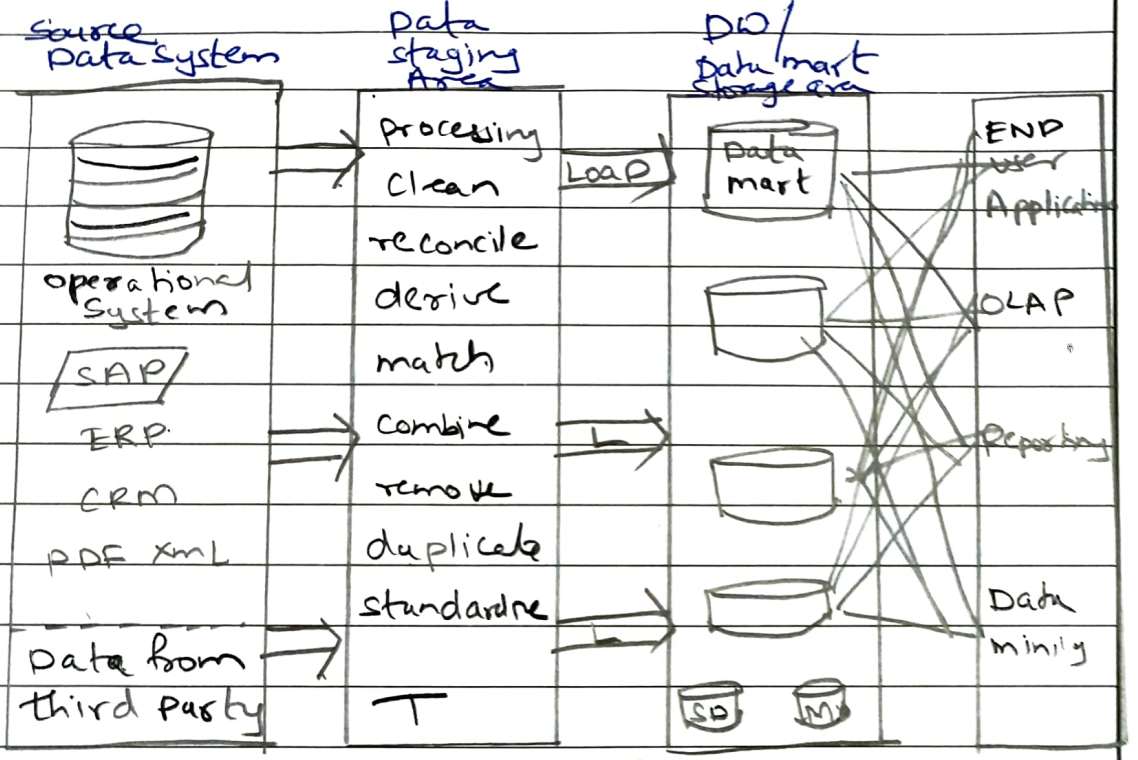



Dendogram.

| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

| Space for Marks | Question No. | START WRITING HERE |
|---|---|---|

Q5 b) Explain three tier data warehouse architecture with a suitable block diagram.




Architecture of DW Systems.

Data Source Component

A DW system gathers info from various data sources. There are mainly 4 types of primary source

1) production data.
2) Internal data
3) External data
4) Archived data.

2) Data staging Component.

It is related to the process

| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

| Space for Marks | Question No. | START WRITING HERE |
|---|---|---|

of extraction transformation & loading of data. It is called ETL process

The Extracted data from variety of sources is temporarily stored in staging area in terms of either flat files or staging ~~area in~~ relational DB. It is transformed into suitable form. It is like preprocessing of data. This preprocessed data is then loaded into DW storage area.

There are 2 types of loading

(a) Initial loading

When DW is implemented & data is transferred for the 1st time into DW storage is called initial loading.

(b) Incremental loading

Once DW starts functioning data from reqd sources is extracted transformed and loaded into DW. It is called Incremental loading and it occurs periodically.

(3) Data storage Component.

It is a seperate repository of data. It must organize data

| Total Marks of Question no. | | Examiner | |
|---|---|---|---|
| | | Moderator | |
| | | Re-Assessor | |

| Space for Marks | Question No. | START WRITING HERE |
|---|---|---|
| | | for efficient analysis. so normally it is based on multi-dimensional data model. |
| | | It also stores meta data which is data about data. |
| | | Each data mart contains the same data as EDW but its scope is restricted to single dept. so it is also called departmental data mart. |
| | 4 | Information Delivery Component. |
| | | It is responsible for delivering the data from data mart or EDW to individual users in specified format |
| | | It include various methods of data delivery such as complex queries, multi-dimensional analysis. cross-dimensional analysis adhoc report generation, various OLAP operations etc. |