B, B2
IT

Devansh Shah
1914078

27/05/21

AD ESE

Dshah (1)

Q1) 1) a

2) d

3) c

4) d

5) a

6) c

7) a

8) a

9) d

10) d

Devansh
1914078
Dsheh (2)

Q1) (B)

(A)

OLTP	OLAP
- It stands for online transaction processing	It stands for online analytical processing
- Used in day to day operations like purchasing, inventory, banking, etc.	Used in data analysis and decision making in - data warehouse system
- The database design is application oriented	The database design is subject oriented
- usage is repetitive	usage is adhoc
- access is read/write index/hash on primary key	access is to lots of scans
- data is current, up to date detailed, flat relational isolated	data is historical, summarized multidimensional integrated, consolidated
- metric is transaction throughput	metric is query throughput

Devansh
19114078

(3)

(d) The advantages of hybrid OLAP are :-

1) HOLAP has better accessibility in comparison to both ROLAP and MOLAP models
2) Higher processing ability when related to ROLAP and MOLAP system
3) Cubes are smaller than MOLAP since only precise data is fetched for processing.
4) It has the volume to occupy huge data as it is a relational database in HOLAP.

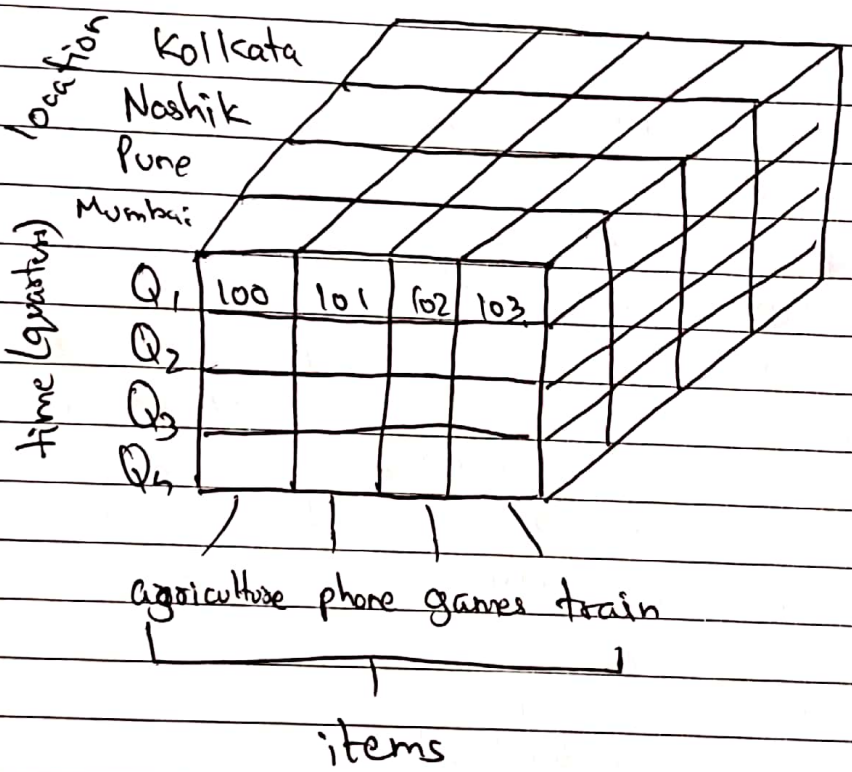
(e) The limitations of ROLAP are :-

1) There are many performance problems associated to the processing of complex queries that require multiple passes through relational data.
2) It is difficult to maintain aggregate tables in data warehouse.
3) Development of middleware to facilitate development of multidimentional applications is required.
4) Does not have complex functions that are provided by OLAP tools.

Devansh
191407
Dzkrah (4)

(c) Multidimensional databases are specialized data stores that organise facts by dimensions, such as geographical region, product line, salesperson, time.

Below is a neat sketch of multidimensional OLAP which represents quantities of a product sold by specific retail locations during certain time periods.

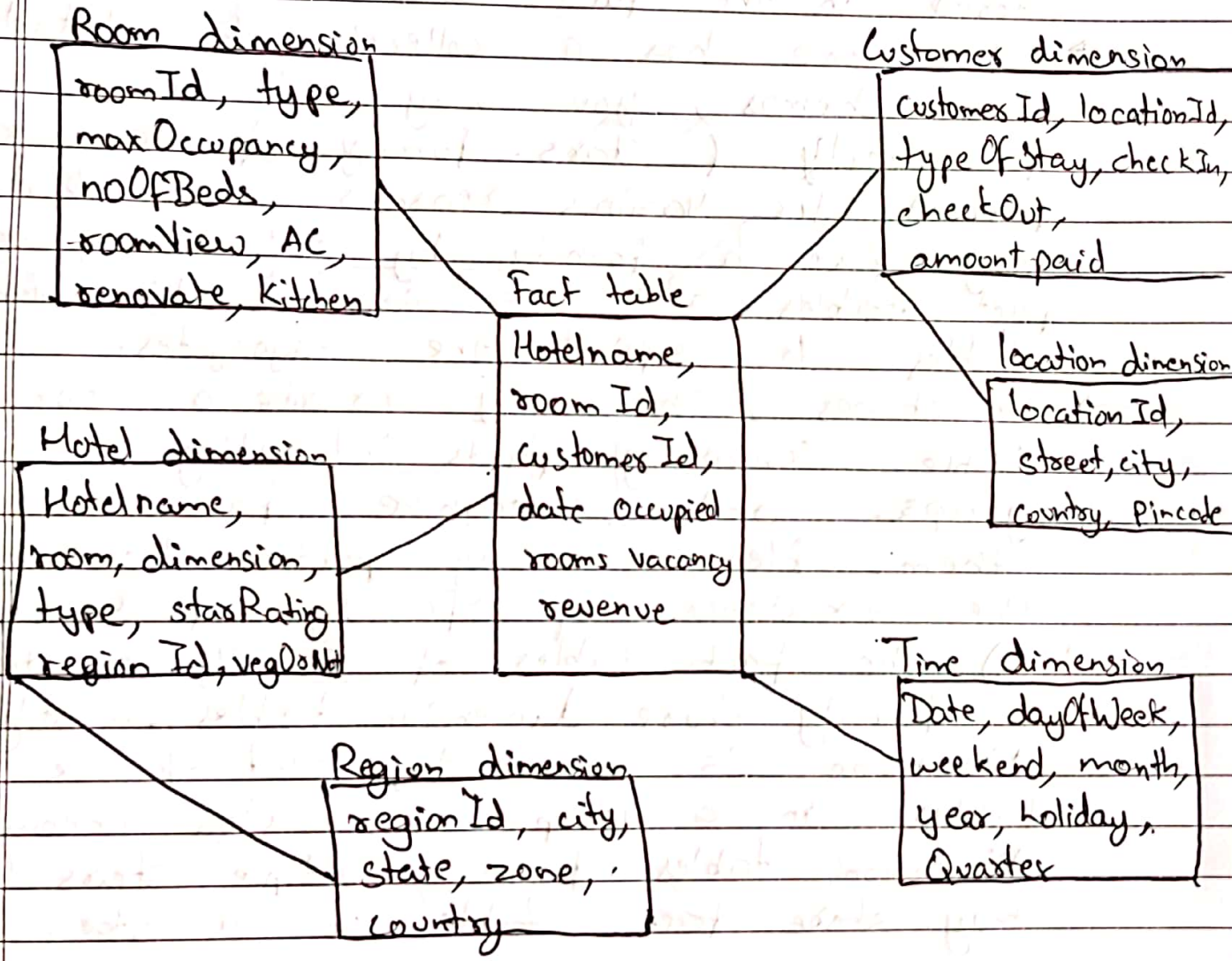


(f) Real world applications of OLAP are:-
- Data mining, marketting
- management reporting, agriculture
- budgeting, forecasting
- Business reporting for sales.

Devansh
1914078
(5)

Q2) i) Snowflake schema for hotel management system

- The schema has one or more normalized dimensions
- Fact table has no. of occupied rooms, no. of vacant rooms and revenue.
- Dimensions are room, time, customer, hotel, region, location



Here, region and location dimension have been normalized.

Dewansh
19140 78
(6) Delhi

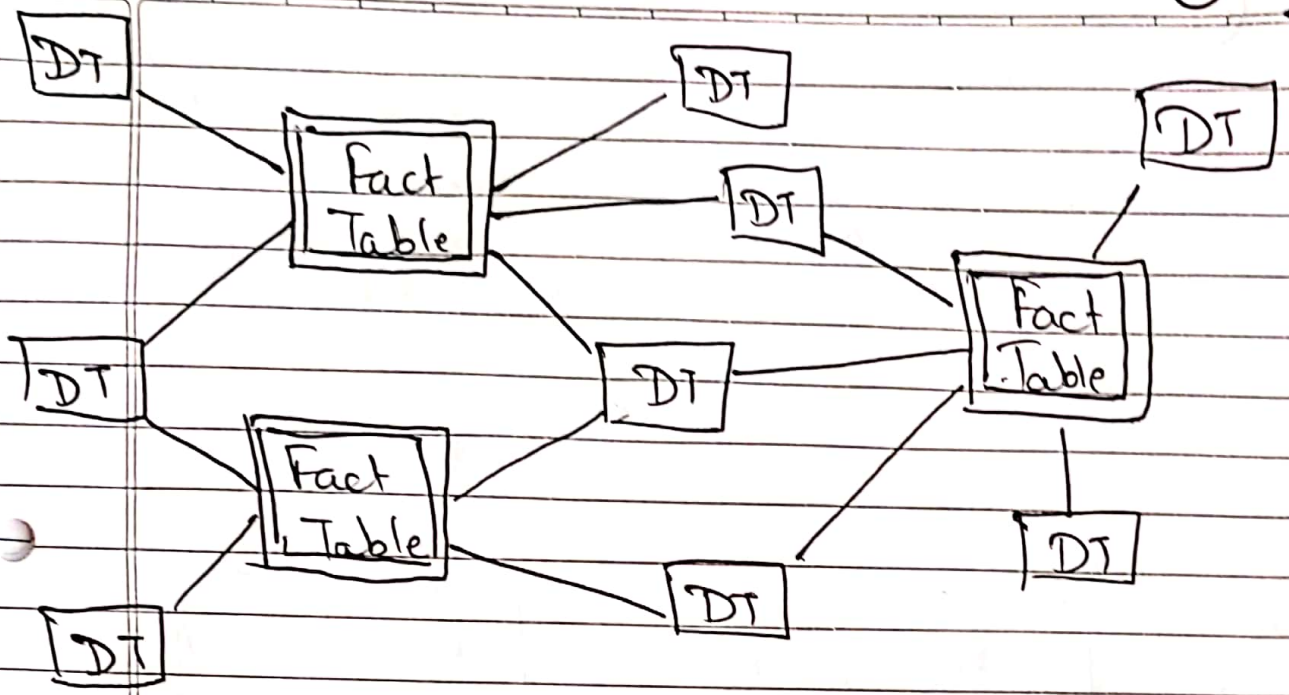
Q2) ii) When you look at single star schema with its fact table and the surrounding dimension tables, you know that it is not the extent of a data warehouse. Almost all the time, all the data warehouse contain multiple star schemas. Each star schema serves a particular purpose to track the measure stored in the fact table.

→ When one has a collection of related star schemas, you may call it a collection of family of stars. Family of stars are formed for various reasons such that one may form a family by just adding aggregate fact tables and the derived dimension tables to support the aggregates.

→ Sometimes, one may create a core fact table containing facts interesting to most users groups and customize fact tables for them. Hence many factors lead to the existence of family of stars.

→ eg) The fact tables of family of stars in a family share dimension tables, mostly time dimension is shared by most of the fact tables in a group. This also means dimension tables form multiple stars that may share fact table of a star.

Devansh
1914078
(7) Dshub



DT : dimension table

Devansh
1914078
(8)

Q3) (i)

Noise removal :-

- Noise is a random error or variance in a measured variable. Noisy data may be due to faulty data collection instruments, data entry problems and technology limitations.
- Multiple ways to handle data are

i) Binning - sort data by consulting the values around it.

eg) Price = 4, 8, 15, 21, 21, 24, 25, 28, 34

Bin a : 4, 8, 15

Bin b : 21, 21, 24

Bin c : 25, 28, 34

(equally partitioned on equal frequency)

ii) Clustering

iii) Smoothing

iv) regression.

(ii)

Data ~~cuts~~ descretization :-

- Data discretization converts a large number of data values into smaller once so that the data evaluation and data management becomes very easy

eg) Age (before discretization) : 10, 11, 17, 19, 38, 40, 42, 70, 73, 80

After discretization.

10, 11, 17, 19 - young, 38, 40, 42 - mature, 70, 73, 78 - old

Devansh
1914078
Delhi

(iii) **Data Smoothing :-**

Data smoothing refers to a statistical approach of eliminating outliers from data sets to make the patterns more noticeable. It's achieved by manipulating data to remove or reduce any volatility or other type of noice.

eg) an economist can smooth out data to make seasonal adjustments for certain indicators like retail sales by reducing the variations that may occur each month like holidays or gas prices.

(iv) **Min-max normalization :-**

In this technique of data normalization, linear transformation is performed on the original data Minimum and maximum value from data is fetched and each value is replaced in accordig to the formula

$$v' = \frac{v - min(A)}{max(A) - min(A)} (newMax(A) - newmin(A)) + newMin(A)$$

v' : new value
v = old value

Devansh
1914078
(10) Dshah

eg) min = 8 marks : 8, 10, 15, 20
max = 20

new Max = 1
new Min = 0

for marks = 8 $V_1 = 8$

$$\text{MinMax} = \frac{V_1 - \text{min marks}}{\text{Max marks} - \text{min marks}} (\text{new max} - \text{new Min}) + \text{new Min}$$

$$= \frac{8-8}{20-8} * (1-0) + 0$$

$$= \frac{0}{12} * 1$$

$$= 0.$$

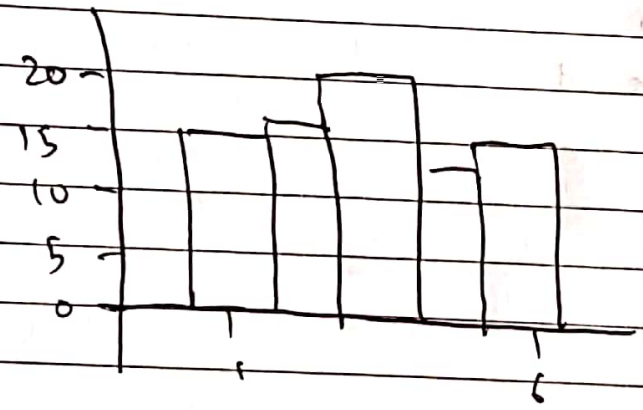
Devansh
19114078
(11) Dshah

v) Z-scores

- Also known as standardized scores, they are scores or data values that have been given a common standard.
- This standard is a mean of zero and a standard deviation of 1.
- Z scores are not necessarily normally distributed.

eg) Scores of 100 people on 2 tests where max of test 1 is 6 and min = 1

It's histogram goes like

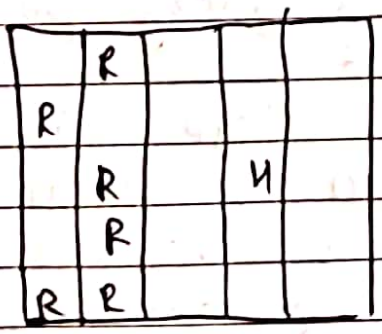


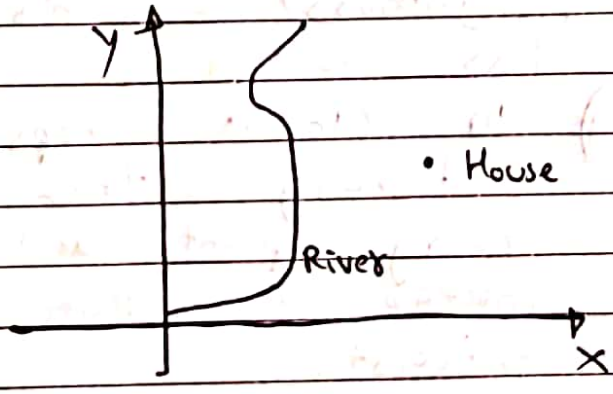
∴ not normally distributed.

Devansh
191408
(12)

Q4) A spatial database is a spatial type of database. It is used for storing and quering of data which are defined in a geometric space such as linear, points or other polygons.
- One can used spatial database to get the view of the real world as we see it, in the form of a database. We can either capture the entire world or a portion of it.
- As we are visualizing the real world so all the measurements input into the database must be as complete and accurate as possible. It should be able to represent real world
- There are peculiar things about spatial databases as they change on the basis of what attributes of the world we capture, when we are capturing them and how much we captured. They are termed as attributes, time period and study sites

Spatial databases are represented in 2 ways :-
Raster model
Vector model





Devansh
M1124078
(13) Dsheb

→ Types of queries :-

1) Range queries :-

- It is used to find all desired selection in a particular range. It's base is a region. Hence also called as ~~relations~~ regional queries.
- Range queries usually need 3 things to be satisfied.
  i) Object of interest
  ii) Query point
  iii) region

eg) Find all the rivers in a 10 km radius to Mumbai city. So, here all the rivers will be objects of interest, Mumbai will be query point, and 10km radius circle with origin at mumbai as region.

2) Nearest Neighbour :-

- The nearest neighbors query, also known as K nearest neighbours (KNN) gives k results as output. It works on the basis of, ~~if~~ it finds out k of the nearest (neighbours) elements to a query point. It has no region constraint.

eg) 10 closest shops to college which sell stationary, then k = 10, college is query point and nearest distance from college to stationary shops are objects of interest.