27/08/2020
Shiva Saran
BE B-20

* Data Mining & Warehousing
* Unit Test 1

- Question 1 (a)

Data preprocessing is a data mining technique which is used to transform the raw data in a useful & efficiant format.

Steps involved in Data pre-processing:

① Data Cleaning

The data can have many irrelevant & missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

② Data transformation

This step is taken inorder to transform the data in appropriate forms suitable for mining process. This involves Normalization, Attribute selection, Discretization & concept hierarchy generation.

③ Data Reduction

Since data mining is a technique that is used to handle huge amount of data. While working with the same, analysis becomes harder in such cases. In order to get rid of this, we use data reduction technique. It aims to increase the storage efficiency & reduce data storage & analysis costs. Variou data reduction steps are:

Data cube aggregation, Attribute subset selection, numerosity reduction & dimensionality reduction.

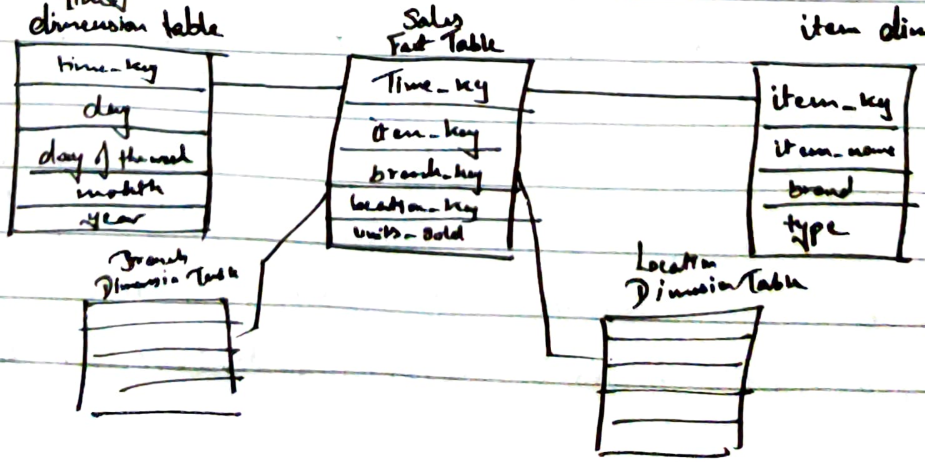
Question 1 (b)

Online Transaction Processing (OLTP)	Online Analytical Processing (OLAP)
(a) Its an online transactional system. It manages database modification.	(a) OLAP is an online analysis & data retrieving process.
(b) Characterised by large numbers of short online transactions.	(b) Characterised by large volume of data.
(c) Uses traditional DBMS.	(c) Uses the data warehouse.
(d) Tables are normalised.	(d) Not normalised - Tables.
(e) Response time is in millisecond.	(e) Response time is in minutes.
(f) Provides fast result for daily used data.	(f) Ensures that response to the query is quicker consistently.

Question 4 (a)

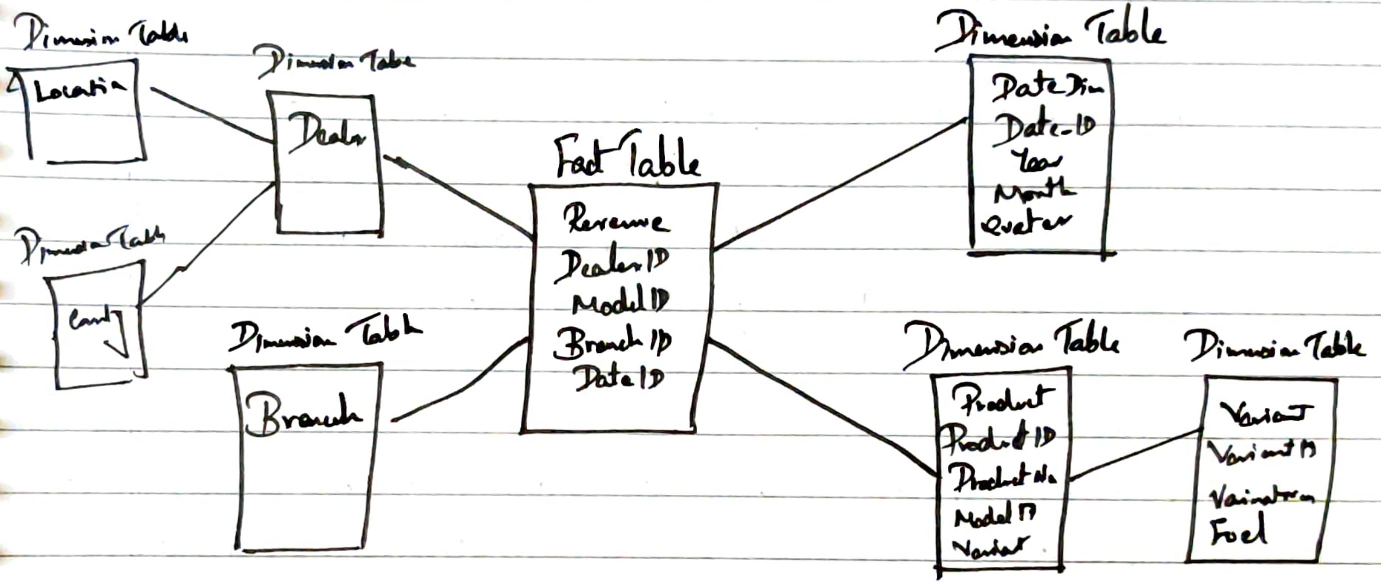
(i) Star Schema

- Each dimension in a star schema is ~~pr~~ represented with one-dimension table.
- The dimension table contains the set of attributes.
- There's a fact table at the center. It contains the keys to each of 4 ~~dimension~~ dimensions.



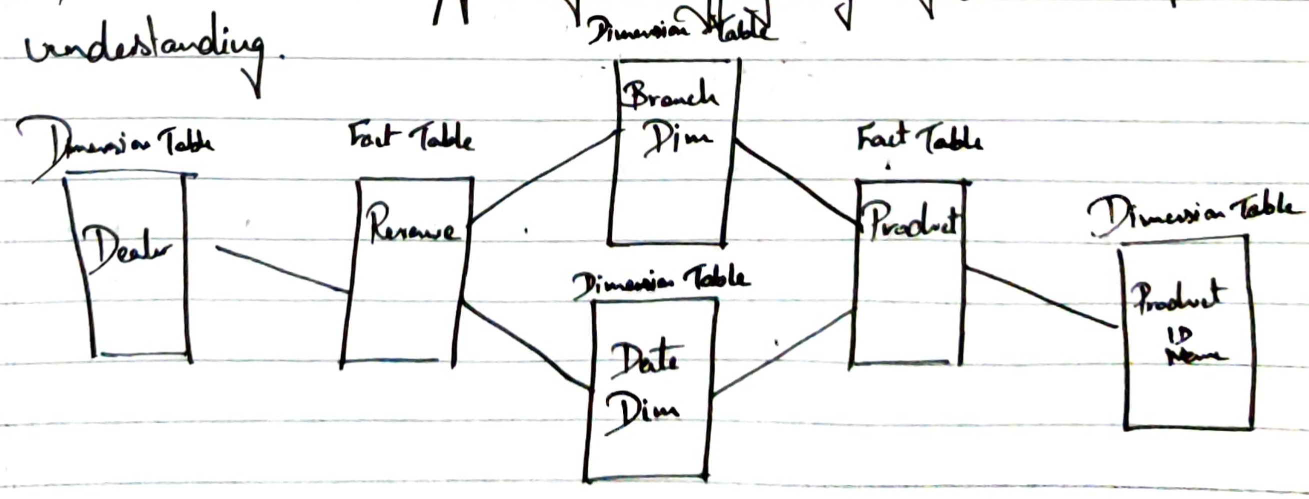
(ii) Snowflake Schema

- Some dimension tables in the snowflake schema are normalised.
- The normalisation splits up the data in ~~d~~ additional ~~ds~~ tables.
- Unlike star schema, the dimension table in a snowflake schema are normalised.



(iii) Fact Constellation Schema

- A fact constellation schema has multiple fact tables. aslo known as galaxy schema.
- It is viewed as a collection of stars.
- This schema is helpful for aggregating fact tables for better understanding.



Question 4 (b)

1. Information Processing

A data warehouse allows to process the data stored in it. The data can be processed by means of querying, basic statistical analysis, reporting using crosstabs, tables, charts or graphs.

2. Analytical Processing

A data warehouse supports analytica processing of the information stored in it. ~~Tg~~ The data can be analysed by means of basic OLAP operations, including slice & dice, drill down, drill up & pivoting.

3. Data mining

Data mining supports knowledge discovery by finding hidden patterns and associations, constructing analytical models, performing classification, & prediction!

Question 6 (a)

1. Minkowski Distance

It is the generalised form of euclidean & Manhattan distance.

$$D = \left(\sum_{i=1}^{n} | p_i - q_i |^p \right)^{1/p}$$

where 'p' is the order parameter.

When p is set to 1, the calculation is same as Manhattan

p is set to 2, it is the same as Euclidean distance.

p is set to ∞, it is chebychev distance.

② Euclidean Distance

It is the straight line distance between 2 data points in a plane.
It is calculated using the minkowski distance formula by setting 'p' value to 2, thus also known as L2 norm distance metric.

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

③ Manhattan Distance

It is the sum of absolute differences between points across all the dimensions.
We get the equation for Manhattan distance by substituting p=1 in the minkowski distance formula.

$$d = \sum_{i=1}^{n} |x_i - y_i|$$

Question 6(b)

Cosine similarity metric is mainly ~~so~~ used to find similarities between two data points. As the cosine distance between the points increases, the cosine similarity or the amount of similarity decreases & vice versa.
Cosine similarity is given by Cos θ & cosine distance is 1 − cos θ.