DATE:

Assignment 1

1) Difference between star and snowflake schema.

	Star schema	Snowflake schema
a.	In star schema the fact table and dimension table are ~~obtained~~ contained.	But in snowflake schema, the fact table, dimension table and sub dimension table are contained.
b.	This is a top-down model.	This is a bottom up model.
c.	Uses more space.	Uses less space.
d.	It takes less time for the execution of queries.	It takes more time for execution of queries.
e.	In star schema, ~~no~~ normalization is not used.	Both normalization and demormalization is used.
f.	Its design is very simple.	Its design is very complex.

2) State any 4 applications of DWH.

Ans i) Business Intelligence - (BI)

Data warehousing is used to support BI tools that helps organizations make data-driven decisions. By consolidating data from different resources into a single repository, we get a unified view of business operations and enables organizations to perform

complex analysis and reporting.

ii) Customer Relationship Management (CRM) -
Data warehousing is used to support CRM applications that help organisations manage their interaction with customers.

iii) Supply chain management -
It is used by integrating data from suppliers, distributions and logistics providers which helps to improve efficiency and identify bottlenecks.

iv) Healthcare -
It is used in applications such as health management, clinical research, patient analytics etc.

5) What is a five number summary of data? Explain with example.

Ans The five number summary is a statistical summary of a dataset. It consists of five values: minimum value, first quartile (Q1), median (Q2), third quartile (Q3), maximum value. It is often used in boxplots and other visualization to show

eg 5, 8, 12, 16, 18, 20, 22, 25, 30, 35

∴ min = 5

Max = 35

median = 19

Q1 = 12

Q3 = 25

DATE:

4) Explain various measures of central tendency of data.

Ans: i. Mean-

The mean is the most common measure of central tendency. It is calculated by adding up all the values in a dataset and then dividing by the total number of values. The mean can be sensitive to extreme values or outliers in dataset.

ii Median-

The median is the middle value of the dataset when the values are arranged in order. It is less affected by outliers unlike mean.

iii Mode-

The mode is the value that appears the most in the dataset. It is useful for describing the most common value in the dataset.

5) Normalize the following group of data: 200, 300, 400, 600, 1000. Use min max normalization by setting new min=0 and max=1.

min = 200 max = 1000.

$$x_{new} = \frac{(x - min)}{(max - min)}$$

$$\therefore x_{new} = \frac{200 - 200}{1000 - 200} = 0$$

$$x_{new} = \frac{300 - 200}{1000 - 200} = 0.125$$

$$x_{new} = \frac{400 - 200}{1000 - 200} = 0.25$$

$$x_{new} = \frac{600 - 200}{1000 - 200} = 0.5$$

$$x_{new} = \frac{1000 - 200}{1000 - 200} = 1$$

0, 0.125, 0.25, 0.5, 1

DATE:

6) What is concept hierarchy? State examples of partial order and total order hierarchy.

Ans a) Concept hierarchy is a way of organizing data or information in a hierarchial structure based on their level of abstraction or generalization.

b) A partial order hierarchy is a ~~way~~ hierarchy in which not all elements are comparable i.e some elements may be incomparable to others.

c) For instance, a eagle is more similar to a sparrow that it is to a shark.

d) A total order hierarchy on the other hand, is a hierarchy in which all elements are comparable to each other.

e) A good example of total order hierarchy is, all integers can be ordered from smallest to largest and every integer can be compared with every other integer.

7) What are different OLAP operations? Explain with examples.

Ans OLAP operations are used to analyze data from different perspectives.

i) Slice -

A slice operation is used to extract information from a cube by selecting single value from its dimensions. For example, to extract data from a specific year from a sales cube, we can slice the cube by the year dimension and select the specific year we are interested in.

DATE:

ii) Dice -
A dice operation is used to extract data from a cube by selecting more than one value from multiple dimensions. For example, for data from specific year and specific product, we select the 'year' and 'product' dimension we are interested in.

iii) Roll-up -
A roll up operation is used to aggregate data from a lower dimension to a higher dimension. For example, to aggregate sales data for a product by year, we can roll up the product dimension by year dimension.

iv) Drill-down -
A drill down operation is used to break down aggregated data from a higher-level dimension to a lower-level dimension. Fo example, to break down sales data ~~from~~ for all products by year, we can drill down the 'product' dimension to the 'year' dimension.

8) Explain in brief major ~~d~~ tasks of data preprocessing.
Ans Data preprocessing is a crucial step in data analysis and machine learning pipelines. It involves cleaning, transforming and preparing ~~new~~ raw data to make it suitable for analysis.

i) Data cleaning -
This involves handling missing values, outliers and inconsistent data.

DATE:

ii) Data integration-
This involves combining data from multiple sources into a single dataset. For example, merging two dataset that have a common key.

iii) Data Transformation-
This involves converting data into a suitable format for analysis. This may involve scaling or normalisation, encoding categorical values.

iv) Data Reduction-
This involves reducing the size of the dataset while retaining as much information as possible. This may involve sampling, feature selection or dimensionality reduction techniques.