Assignment : Data Cleaning, Transformation and Modelling

Question 1 : What is Data Loading in Power BI and why is it considered the first step of analysis?

Solution : Data Loading in Microsoft Power BI refers to the process of importing data from different data sources into Power BI so that it can be transformed, modeled, and analyzed.

It is done using the Get Data option in Power BI Desktop.

Power BI allows loading data from:

- Excel files
- CSV/Text files
- SQL Server databases
- Web sources
- Cloud platforms (Azure, SharePoint, etc.)
- APIs and online services

Once data is loaded, it becomes available in:

- **Data View** (to see tables)
- **Model View** (to create relationships)
- **Report View** (to build dashboards and visuals)

Why is Data Loading Considered the First Step of Analysis?

Data loading is the foundation of the entire BI process. Without data, no analysis is possible.

1 :it Brings Raw Data into the System

2 :Enables Data Cleaning & Transformation

3 : Forms the Base for Data Modeling

4 :Ensures Accurate Insights

5 :Connects Multiple Data Sources

Question 2 : Explain the difference between “Load” and “Transform Data” in Power BI.

Solution :

Load

- Imports data directly into Power BI.
- No changes are made.
- Used when data is already clean.

Transform Data

- Opens Power Query Editor.
- Allows cleaning and modifying data (remove duplicates, change data types, filter rows).
- Data is loaded only after applying changes.

Basis	Load	Transform Data
Purpose	Directly import data	Clean and modify data before importing
Data Changes	No changes	Data can be cleaned and reshaped
Tool Used	Direct load	Power Query Editor
Best For	Clean datasets	Raw or messy datasets
Risk Level	May load incorrect data	Improves data accuracy

Question 3 : What is a Fact Table and a Dimension Table? Give examples from the dataset.

Solution :

Fact Table

A Fact Table contains measurable, numerical data used for analysis.

Characteristics:

- Contains numbers (sales, price, quantity, revenue)
- Large number of records
- Connected to dimension tables using foreign keys

Example (from a sales/property dataset):

Property_ID	Price	Area_sqft	Bedrooms

Here:

- **Price**
- **Area_sqft**

- **Bedrooms**

These are measurable values → so this table acts as a **Fact Table**

Dimension Table

A Dimension Table contains descriptive information about the facts.

Characteristics:

- Contains text or descriptive fields
- Provides context to fact data
- Used for filtering and grouping

Example:

Property_ID	City	Location_Details	Owner_Name

Here:

- **City**
- **Location_Details**
- **Owner_Name**

These describe the property → so this is a Dimension Table.

Question 4 : Why is Star Schema preferred over Snowflake Schema in Power BI?

Solution :

1 :Simpler Structure

- **Star Schema**: One central Fact table connected directly to Dimension tables.
- **Snowflake Schema**: Dimension tables are further divided into multiple related tables.

Star Schema is easier to understand and design.

2: Better Performance

- Star Schema has **fewer joins** between tables.
- Fewer joins = **faster query performance** in Power BI.

Snowflake Schema requires more joins, which can slow down reports.

3: Easier DAX Calculations

In Star Schema:

- Relationships are simple.
- Writing DAX measures becomes easier.
- Fewer chances of ambiguity.

Snowflake Schema can create complex relationship paths.

4: Better for Reporting & Visualization

Power BI works best with:

- Clear fact table (numbers)
- Clear dimension tables (descriptions)

Star Schema supports this structure perfectly.

Practical Questions

Question 5 : Identify and remove duplicate records based on Date, Country, and State.

Solution :

Sample Data (Before Removing Duplicates)

Date	Country	State	Confirmed	Recovered	Deaths
01-04-2020	India	Maharashtra	5000	400	200
01-04-2020	India	Maharashtra	5000	400	200
01-04-2020	USA	California	8000	1000	300
02-04-2020	India	Maharashtra	5500	600	220

The first two rows are duplicates (same Date + Country + State).

Steps in Power BI

1. Home → Transform Data
2. Select Date, Country, State
3. Remove Rows → Remove Duplicates
4. Close & Apply

Data After Removing Duplicates

Date	Country	State	Confirmed	Recovered	Deaths
01-04-2020	India	Maharashtra	5000	400	200
01-04-2020	USA	California	8000	1000	300
02-04-2020	India	Maharashtra	5500	600	220

Duplicate row removed successfully.

Question 6 : Identify and replace null values in vaccination-status.

Solution : **Sample Data (Before Handling Nulls)**

Date	Country	State	Confirmed	Recovered	Deaths
03-04-2020	India	Delhi	2000	null	50
03-04-2020	USA	null	9000	1200	350

Problems:

- Recovered has null value
- State has null value

Steps in Power BI

1. Transform Data
2. Select column
3. Transform → Replace Values
4. Replace:
    a. null (numeric) → 0
    b. null (text) → "Unknown"

Data After Replacing Nulls

Date	Country	State	Confirmed	Recovered	Deaths
03-04-2020	India	Delhi	2000	0	50
03-04-2020	USA	Unknown	9000	1200	350

Null values handled properly.

Question 7 : Create a new column to calculate Recovery Rate.

Solution :

FORMULA

Recovery Rate = (Recovered / Confirmed) × 100

DAX Formula

```
Recovery Rate =
DIVIDE(
    Corona_Virus_2020[Recovered],
    Corona_Virus_2020[Confirmed],
    0
) * 100
```

Sample Output

Country	Confirmed	Recovered	Recovery Rate (%)
India	5000	400	8%
USA	8000	1000	12.5%

Calculation Example:
India → (400 / 5000) × 100 = 8%

New calculated column created successfully.

Question 8 : Create a summarized table showing total confirmed cases by Country

Solution :

Original Data

Country	Confirmed
India	5000
India	5500
USA	8000
USA	9000

DAX Formula

```
Country_Summary =
SUMMARIZE(
    Corona_Virus_2020,
    Corona_Virus_2020[Country],
    "Total Confirmed Cases",
    SUM(Corona_Virus_2020[Confirmed])
)
```

Summarized Output Table

Country	Total Confirmed Cases
India	10500
USA	17000

India Total = 5000 + 5500 = 10500

USA Total = 8000 + 9000 = 17000