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

Solution:

Introduction Data loading is the process of bringing raw information from various external sources—like Excel files, SQL databases, or web pages—into the Power BI environment. Before we can create any charts or find trends, the data must actually exist within the tool so it can be processed.

Main Explanation Data loading is considered the foundational first step of analysis for several reasons:

- **Foundation of the Model:** Without loading data, there is no "raw material" to work with. It is the bridge between where the data lives (the source) and where the insights are created (Power BI).
- **Defining the Scope:** When you load data, you are essentially defining the boundaries of your analysis. You decide which tables and which columns are relevant to the problem you are trying to solve.
- **Data Connectivity:** Power BI establishes a connection during this phase. This connection allows the software to understand the format and structure of the source, making it possible to refresh the data later when the source file changes.
- **Memory Management:** Power BI uses an in-memory engine (VertiPaq). Loading the data compresses it and stores it in the system's memory, which is why Power BI can perform complex calculations so quickly.

Final Conclusion In short, data loading is the "entry point" of the entire analytical workflow. You cannot transform, model, or visualize data that hasn't been brought into the system first.

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

Solution:

Introduction When you connect to a data source in Power BI, you are usually presented with two main options: "Load" and "Transform Data." Choosing between them determines whether your data goes straight into the report or takes a detour through an editor for cleaning.

Main Explanation The differences between the two are based on the "health" and readiness of your data:

- **Load:**
  - This option brings the data directly into the Power BI Data Model exactly as it exists in the source.
  - It is used when the data is already "clean"—meaning it has correct headers, no empty rows, and proper data types.
  - It is the faster route if no preparation is required.

- **Transform Data:**
  - This option opens the **Power Query Editor**.
  - It is used when the data is "messy." You might need to remove duplicate rows, change text to numbers, split columns (like "First Name" and "Last Name"), or filter out unnecessary information.
  - It allows you to create a "recipe" of steps that Power BI will follow every time the data is refreshed to ensure it stays clean.

Final Conclusion The main difference is that "Load" is for ready-to-use data, while "Transform Data" is for data that requires cleaning and restructuring before it can be used for accurate reporting.

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

Solution:

Introduction In data modeling, we organize information into two specific types of tables to make analysis easier: Fact tables and Dimension tables. This structure helps the computer process large amounts of data efficiently.

Main Explanation

- **Fact Table:**
  - This is the central table that contains the quantitative data or "numbers" we want to calculate (sums, averages, etc.).
  - It usually contains a large number of rows and tracks specific "events" or "transactions."
  - *Example from the dataset:* A **Sales Table** is a perfect fact table because it contains numbers like "Quantity Sold," "Price," and "Discount."

- **Dimension Table:**
  - These are the "lookup" tables that provide context to the numbers in the fact table. They describe the "who, what, where, and when."
  - They contain descriptive attributes (text).
  - *Example from the dataset:* A **Product Table** (containing Product Name and Category) or a **Customer Table** (containing Customer Name and Location) are dimension tables. They tell us *which* product was sold in the Sales Table.

Final Conclusion Fact tables provide the "numbers" (the What), while Dimension tables provide the "details" (the Context). Together, they allow us to see not just *how much* we sold, but *which* items were most popular.

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

Solution:

Introduction The Star Schema and Snowflake Schema are two ways of organizing tables in a data model. While both are used, the Star Schema is widely considered the "Gold Standard" for Power BI because of its simplicity and speed.

Main Explanation The preference for Star Schema is based on three main factors:

- **Performance (Speed):** In a Star Schema, the Fact table is directly connected to Dimension tables. This means Power BI only has to make one "jump" (join) to filter data. In a Snowflake Schema, dimensions are broken into sub-dimensions (normalized), requiring multiple jumps, which slows down the report.
- **Simplicity for Users:** The Star Schema is much easier to understand. When a user looks at the "Fields" pane, they see a clear list of categories. In a Snowflake Schema, the list of tables becomes long and confusing, making it harder to build visuals.
- **DAX Efficiency:** Formulas written in DAX (Data Analysis Expressions) perform better and are easier to write when the model is simple. Complex relationships in a Snowflake Schema can sometimes lead to unexpected results or circular dependency errors.

Final Conclusion Power BI is optimized for the Star Schema. By keeping the model "flat" and reducing the number of relationships between tables, the reports run faster and the data remains much easier to manage.

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

Solution

Introduction

Duplicate records can distort analysis by inflating totals and creating misleading results. In this dataset, a duplicate means that the same **Date**, **Country**, and **State** combination appears more than once. Removing such records helps maintain data accuracy.
Explanation (Power BI Method)

This task is performed using **Power Query Editor** in Power BI:

1. Click **Transform Data** to open Power Query Editor.
2. Select the **Date**, **Country**, and **State** columns together (hold the Ctrl key while selecting).
3. Right-click on any selected column header.
4. Click **Remove Duplicates**.
5. Power BI keeps only unique combinations and removes repeated rows.
6. The step is saved in **Applied Steps**, so it updates automatically on data refresh.

Conclusion
After removing duplicates, each state's data for a given date is counted only once, ensuring reliable totals and accurate reporting.

Question 6: Identify and replace null values in Vaccination_Status

Solution

Introduction

Null or blank values can create confusion in reports and slicers. Instead of leaving vaccination status empty, replacing null values with a meaningful label improves clarity and consistency.

Explanation (Power BI Method)

1. Open **Power Query Editor**.
2. Select the **Vaccination_Status** column.
3. Right-click and choose **Replace Values**.
4. In the dialog box:
    - **Value to find:** Leave blank (this targets null values).
    - **Replace with:** Enter Unknown.
5. Click **OK**.

Conclusion
Replacing null values ensures that all records appear correctly in visuals and filters, without showing blank categories.

Question 7: Create a new column to calculate Recovery Rate

Solution

Introduction
The **Recovery Rate** shows how many confirmed patients have recovered, relative to total confirmed cases. This metric helps compare recovery performance across different regions.

Explanation (Power BI Method)

1. In **Power Query Editor**, go to the **Add Column** tab.
2. Click **Custom Column**.
3. Name the column Recovery Rate.
4. Enter the formula:
5. [Recovered_Cases] / [Confirmed_Cases]
6. Click **OK**.
7. Change the data type to **Decimal Number** or **Percentage**.
8. To avoid errors when confirmed cases are zero, use this safer formula:
9. if [Confirmed_Cases] = 0 then 0 else [Recovered_Cases] / [Confirmed_Cases]

Conclusion
The Recovery Rate column provides deeper insight than raw numbers and helps analyze trends more effectively.

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

Solution

Introduction

To analyze the overall impact of COVID-19 at the country level, it is useful to aggregate all state-level records into a single total per country.

Explanation (Power BI Method)

1. Open **Power Query Editor**.
2. Select the **Country** column.
3. Click **Group By** from the Home tab.
4. In the Group By window:
    - **Group by:** Country
    - **New column name:** Total Confirmed Cases
    - **Operation:** Sum
    - **Column:** Confirmed_Cases
5. Click **OK**.

Conclusion

The summarized table makes it easy to create high-level visuals such as bar charts or maps, allowing quick comparison of total confirmed cases across countries.