USN NUMBER: 1RVU23CSE264
NAME: Meghana G

| Ex No: 1<br>Date: 6/8/25 | Exploring the Data Engineering Lifecycle and Stakeholder Roles |
|---|---|

## Objective:

This lab provides hands-on experience exploring the data engineering lifecycle and understanding the roles of key stakeholders. Participants will simulate responsibilities of data engineers, data scientists, and business analysts while examining raw data sources and planning a data-driven solution.

## Outcomes:

1. Identify and describe each stage of the data engineering lifecycle.
2. Explain the specific responsibilities of stakeholders across the lifecycle.
3. Collaborate to define a business problem using raw data sources.
4. Draft a requirements document based on the business use case.

## Materials:

- Raw sales data CSV file (`sales_data_raw.csv`)
- Customer feedback JSON file (`customer_feedback.json`)
- Folder structure representing a mock data warehouse or data lake

## Lab Procedure:

### Stage 1: Problem Definition and Requirements Gathering (Business Analyst)

1. Review both datasets provided (`sales_data_raw.csv` and `customer_feedback.json`).

- Sales_price.csv contains:

  `product_id`, `sale_price`, `quantity`, `customer_id`, and `sale_date`.

- Needs cleaning: missing values and $ symbol in sale_price.

- `customer_feedback.json` contains:

  `product_id`, `customer_id`, `sentiment_score`, and `review_date`.

USN NUMBER: 1RVU23CSE264
NAME: Meghana G

2. Formulate a business question, e.g., "What are the top 5 products by revenue, and how does customer sentiment vary for them?"
3. Identify required data points (e.g., `product_id`, `sale_price`, `customer_id`, `sentiment_score`).

| Field | Source | Purpose |
|---|---|---|
| `product_id` | Both files | Common key for merging |
| `sale_price` | sale_price.csv | To calculate revenue |
| `quantity` | sale_price.csv | To calculate revenue |
| `sentiment_score` | customer_feedback.json | To analyze customer perception |
| `customer_id` | Optional (for filtering) | Track individual-level trends |

4. Create a short requirements document outlining the problem, key metrics, and desired insights.

   **Problem**: Identify high-revenue products and assess if they align with positive customer feedback.

   **Key metrics:** Total Revenue per Product, Average Sentiment Score per Product

   **Desired insights:** Top 5 revenue products, Sentiment distribution for those products

**Stage 2: Role-Based Collaboration Simulation**

1. Discuss and map out how the Data Engineer will ingest and clean the data.

- **Ingest:** Load CSV and JSON files into Pandas DataFrames.
- **Clean:** Strip `$` from `sale_price`, convert to float, Handle missing values in quantity, Standardize date formats.
- After done with ingest and clean need to save cleaned data to `/processed/cleaned_sales.csv` and `/processed/cleaned_feedback.csv`

2. Identify how the Data Scientist will analyze and model insights based on the cleaned data.

- Merge datasets on `product_id`.
- Compute:
  - Total revenue = `sale_price × quantity`
  - Average sentiment per product.
- Analyze:
  - Rank products by revenue.

  - Compare revenue vs sentiment using visualizations (bar/box plots).
- Output: Summary table of top 5 products with revenue and average sentiment.

3. Define how the Business Analyst will interpret and report results.

   **Interpret:**

   - Check if top-selling products have good/bad sentiment.
   - Identify gaps between revenue and satisfaction.

   **Report:**

   - Create visual and textual report (e.g., PowerPoint or PDF).
   - Recommend actions (e.g., improve low-sentiment high-revenue items).

4. Define how each stakeholder contributes to the overall data solution.

| Stakeholder | Contribution |
|---|---|
| **Business Analyst** | Define problem, metrics, and communicate insights |
| **Data Engineer** | Ensure data integrity, availability, and cleanliness |
| **Data Scientist** | Extract insights, model trends, create visualizations |

5. Document the flow of responsibilities and dependencies between roles.

   Business Analyst (Defines Problem) --> Data Engineer (Cleans Data) --> Data Scientist (Analyzes & Models) --> Business Analyst (Interprets & Reports).

**Conclusion:** This lab highlights the collaborative nature of data-driven projects. By defining clear roles and responsibilities across the lifecycle, teams can efficiently transform raw data into actionable insights that support strategic business decisions.