USN NUMBER: 1RVU23CSE227
NAME: KRUTHIK S

Ex No: 1 Date: 06-08-205	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`).

   sale_price.csv: Contains sale_id, product_id, customer_id, sale_price, quantity, sale_date.

   customer_feedback.json: Contains product_id, customer_id, sentiment_score, review_text, review_date.

USN NUMBER: 1RVU23CSE227
NAME: KRUTHIK S

2. Formulate business question:

   "**Which region is projected to have the highest customer traffic in the upcoming holiday season, and how does the average customer sentiment compare across regions**?"

3. Identify required data points

   From **sale_price.csv**:

   - store_id or region - to group sales by location
   - sale_date - to filter for recent months or holiday season period
   - customer_id - to count unique customers

   From **customer_feedback.json**:

   - region or store_id - to align sentiment with traffic data
   - sentiment_score - to compare satisfaction levels by region

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

   Business Problem:

   Determine which region is likely to experience the highest customer traffic in the upcoming holiday season and compare average customer sentiment across those regions.

   Key Metrics:

   - Customer traffic per region - count of unique customers
   - Average sentiment score per region

   Desired Insights:

   - Identify the top region by predicted holiday traffic
   - Compare satisfaction levels between high-traffic and low-traffic regions
   - Provide recommendations for resource allocation and marketing focus

USN NUMBER: 1RVU23CSE227
NAME: KRUTHIK S

Stage 2: Role-Based Collaboration Simulation

1. Data Engineer - Ingest & Clean Data

- Load sales and feedback data into Pandas.
- Parse sale_date and filter for relevant date range .
- Ensure region or store_id field is clean and consistent between both datasets.
- Remove duplicate customer_id entries per region for accurate traffic counts.
- Clean sentiment scores and handle missing values.
- Export cleaned datasets.

2. Data Scientist - Analyze & Model Insights

- Count **unique customers per region** to measure traffic.
- Use past seasonal data to project holiday traffic (simple trend projection or time-series forecast if available).
- Calculate **average sentiment per region** from feedback data.
- Merge traffic and sentiment results.
- Visualize:
    - Bar chart of projected traffic by region.
    - Side-by-side sentiment comparison.

3. Business Analyst - Interpret & Report Results

- Identify regions with high traffic but low sentiment - target for service improvement.
- Identify high traffic & high sentiment regions - focus marketing & promotions.
- Recommend staffing, stock, and marketing adjustments for the holiday period.

4.Stakeholder Roles & Contributions

Business Analyst (BA)	- Defines the problem in the context of the holiday season. - Identifies required data points: region, store ID, customer IDs, sale dates, sentiment scores. - Specifies KPIs: projected customer traffic and average sentiment per region. - Communicates these requirements to the technical team.

Data Engineer (DE)	- Ingests sales and feedback data from CSV and JSON sources. - Cleans sale_date and region fields for consistency.

	- Ensures customer IDs are unique per region for traffic counts. - Outputs a clean, structured dataset linking sales with customer sentiment by region.

Data Scientist (DS)	- Uses historical sales patterns to project holiday traffic per region. - Calculates unique customer counts (traffic) per region. - Computes average sentiment scores per region from feedback data. - Produces visualizations showing traffic and sentiment trends.

Business Analyst (BA) (end stage)	- Interprets analytical results in the business context. - Identifies regions with high traffic but low sentiment for improvement strategies. - Prepares final recommendations for staffing, inventory, and marketing focus.

5. Flow of Responsibilities & Dependencies

[Business Analyst] -> Defines problem, KPIs

↓

[Data Engineer] -> Ingests, cleans, and structures data

↓

[Data Scientist] -> Analyzes data, generates insights

↓

[Business Analyst] -> Interprets and reports results

GitHub Link: https://github.com/kruth-s