USN NUMBER: 1RVU23CSE232
NAME: KUSHAL U

Ex No: 1 Date: 13-08-2025	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`).
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`).
4. Create a short requirements document outlining the problem, key metrics, and desired insights.

Stage 2: Role-Based Collaboration Simulation

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

USN NUMBER: 1RVU23CSE232
NAME: KUSHAL U

2. Identify how the Data Scientist will analyze and model insights based on the cleaned data.
3. Define how the Business Analyst will interpret and report results.
4. Define how each stakeholder contributes to the overall data solution.
5. Document the flow of responsibilities and dependencies between roles.

Stage 1:

2.Formulate a business question?

1. Which products have the highest customer satisfaction, and do they align with the top revenue-generating products?
2. How does customer sentiment vary by product category, and what impact does this have on sales performance?
3. Which products have high sales volume but low customer sentiment, indicating potential quality issues?
4. What is the relationship between average sale price and customer sentiment for each product?
5. Do products with increasing sentiment over time also show an increase in sales revenue?

3.Identify required data points?

From sales_data_raw.csv

- product_id
- sale_price
- quantity
- sale_date

From customer_feedback.json

- product_id
- sentiment_score
- customer_id
- review_date

USN NUMBER: 1RVU23CSE232
NAME: KUSHAL U

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

Problem Statement

The business wants to understand product performance not just in terms of revenue, but also customer satisfaction. By analyzing sales data alongside customer feedback, we aim to identify products that are both high revenue generators and well-received by customers, as well as products that may need improvement.

Key Metrics

1. Total Revenue per Product = sale_price × quantity
2. Average Sentiment Score per Product = mean of sentiment_score
3. Top Products by Revenue – ranked list based on total revenue
4. Sentiment Distribution – range and variance of sentiment scores for top products

Desired Insights

- Identify top 5 products by revenue.
- Understand average customer sentiment for each top product.
- Spot mismatches where products have high revenue but low sentiment.
- Recommend focus areas for marketing, quality improvement, or inventory planning based on findings.

Stage 2: Role-Based Collaboration Simulation

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

Data Ingestion & Cleaning Plan

1. Data Ingestion

Goal: Bring raw data from both sources into a unified, analyzable format.

Steps:

1. Source Identification
    - sales_data_raw.csv → Sales transactions data (possibly from POS or ERP system).
    - customer_feedback.json → Customer sentiment data (possibly from review platform or CRM).

2. Data Import

    - Use ETL pipeline or scripting (Python + Pandas, or Apache Spark for scale).
    - Load CSV using pd.read_csv() and JSON using pd.read_json() or equivalent Spark functions.

3. Schema Definition

    - Ensure both datasets have well-defined column names and consistent data types.
    - Map product_id as the primary key for joins.

2. Data Cleaning

Goal: Remove inconsistencies, handle missing values, and standardize formats.

Sales Data (sales_data_raw.csv)

- Price Cleaning: Remove $ symbols and convert sale_price to float.
- Quantity Cleaning: Replace missing quantity values with 0 or impute based on historical averages.
- Date Standardization: Convert sale_date to YYYY-MM-DD format.
- Type Casting: Ensure product_id and customer_id are strings for consistency.

Customer Feedback Data (customer_feedback.json)

- Date Standardization: Convert review_date to YYYY-MM-DD. Handle mixed formats (e.g., YYYY/MM/DD, MM/DD/YYYY).
- Type Casting: Ensure sentiment_score is float.
- Duplicates Handling: Remove exact duplicate rows (same customer_id, product_id, and date).
- Validation: Drop or flag sentiment scores outside the valid range (e.g., 0–5).

3. Data Integration

- Join datasets on product_id (and customer_id if needed for more granular analysis).
- Ensure the join preserves all relevant transactions, handling products with no feedback and feedback with no sales data.

USN NUMBER: 1RVU23CSE232
NAME: KUSHAL U

4. Storage & Access

- Store cleaned, merged data in a central data warehouse (e.g.AWS Redshift).
- Save intermediate cleaned datasets in a data lake (e.g., AWS S3) for reprocessing if needed.

5. Automation

- Schedule ingestion and cleaning via Airflow or Prefect to run daily/weekly.
- Implement data validation checks (schema enforcement, null value thresholds) before loading into analytics systems.

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

Data Scientist Analysis & Modeling Plan

1. Data Exploration (EDA)

- Review sales trends (total revenue, quantity sold per product).
- Explore sentiment distribution across products.
- Identify outliers in pricing, sales volume, or sentiment scores.

2. Key Metrics Calculation

- Total Revenue per Product = `sale_price × quantity`.
- Average Sentiment per Product = mean(`sentiment_score`).
- Rank products by revenue and sentiment score.

3. Comparative Analysis

- Compare top revenue products with highest sentiment scores.
- Flag products with high revenue but low sentiment (potential quality issues).
- Analyze correlation between price and sentiment.

4. Modeling Insights

- Build predictive models (optional) to forecast revenue based on sentiment trends.
- Use regression analysis to see if sentiment scores significantly impact sales.
- Apply clustering to segment products into groups (high sales–high sentiment, high sales–low sentiment, etc.).

USN NUMBER: 1RVU23CSE232
NAME: KUSHAL U

5. Visualization & Reporting

- Create bar charts for revenue vs sentiment.
- Use scatter plots to show relationships between price and sentiment.
- Build dashboard (Tableau, Power BI, or Python/Plotly) for ongoing monitoring
- Define how the Business Analyst will interpret and report results.

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

Business Analyst – Interpretation & Reporting

Interpretation:

- Find top products by sales and see how customers feel about them.
- Spot products with high sales but low sentiment (quality issues).
- Spot products with high sentiment but low sales (promotion opportunities).

Reporting:

- Create charts and summaries to show trends clearly.
- Give clear recommendations for marketing, quality, or inventory changes.
- Share results through dashboards, PDF reports, and short presentations**.**

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

Stakeholder Roles in the Data Solution

1. Data Engineer

- Collects (ingests) sales and feedback data.
- Cleans and prepares data for analysis.
- Ensures data is stored and updated in a reliable system.

2. Data Scientist

- Analyzes cleaned data to find trends and patterns.
- Builds models to predict or explain relationships (e.g., sentiment vs sales).
- Creates visualizations to show insights clearly.

3. Business Analyst

- Interprets analysis in a business context.
- Identifies opportunities, risks, and recommendations.
- Communicates results to management and other teams.

USN NUMBER: 1RVU23CSE232
NAME: KUSHAL U

4. Management / Decision Makers

- Reviews the reports and dashboards.
- Decides on marketing, pricing, and quality strategies based on insights.
- Tracks changes and measures impact.

5.Do products with increasing sentiment over time also show an increase in sales revenue?

Flow of Responsibilities & Dependencies

1. Data Engineer → Data Scientist
    - Cleans and prepares the data.
    - Scientists need this clean data to do analysis.

2. Data Scientist → Business Analyst
    - Analyzes data, finds patterns, and makes charts.
    - An analyst needs this to understand and explain results.

3. Business Analyst → Management
    - Turns analysis into business recommendations.
    - Management uses this to make decisions.

4. Management → All Roles
    - Gives direction and feedback.
    - Everyone depends on management for priorities.