# LabReport:Exploring the Data Engineering Lifecycle and Stakeholder Roles

Ex No: 1

Date: 13 August 2025

### 1. Requirements Document (Business Analyst Role)

This document outlines the business problem and data requirements, simulating the first stage of the data engineering lifecycle.

* **Business Problem Formulation:** After reviewing the provided sales\_data\_raw.csv and customer\_feedback.json files, the primary business question is: **“What are the top 5 products by revenue, and how does their customer sentiment compare?”** This question aims to link financial performance directly with customer satisfaction to identify high-value products that are also well-regarded by customers.
* **Key Metrics and Desired Insights:**
  + **Key Metrics:**
    - Total Revenue per Product.
    - Average Sentiment Score per Product.
  + **Desired Insights:** The final output should be a consolidated report that ranks the top 5 products by total revenue and displays their corresponding average sentiment score, allowing for a direct comparison between financial success and customer happiness.
* **Required Data Points:**
  + From sales\_data\_raw.csv: product\_id, sale\_price, quantity.
  + From customer\_feedback.json: product\_id, sentiment\_score.

### 2. Role-Based Collaboration Plan

This section details the simulated responsibilities for each stakeholder, mapping out how the data solution will be developed.

#### Data Engineer Responsibilities

The Data Engineer's primary role is to build and maintain the data architecture, ensuring clean and reliable data is accessible for analysis.

* **Data Ingestion (Extract):**
  + Extract structured sales data from sales\_data\_raw.csv.
  + Extract semi-structured feedback data from customer\_feedback.json.
* **Data Cleaning & Transformation (Transform):**
  + **Cleaning:** Handle missing values (e.g., in sale\_price or quantity) and inconsistent data formats (e.g., in sale\_date).
  + **Normalization:** Standardize the sale\_price column by removing currency symbols ($) and converting all values to a numeric type for accurate calculations.
  + **Feature Engineering:** Calculate a total\_sale for each transaction (sale\_price \* quantity).
  + **Joining:** Combine the cleaned sales data with the customer feedback data into a single, unified table using product\_id as the common key.
* **Data Loading (Load):**
  + Load the final, processed dataset into a mock data warehouse (e.g., a new CSV file in a processed folder), making it ready for analysis.

#### Data Scientist Responsibilities

The Data Scientist uses the curated dataset to perform deep analysis and extract insights.

* **Data Analysis:**
  + Work with the clean, integrated data provided by the Data Engineer.
  + Group the data by product\_id and aggregate the total\_sale to calculate the total revenue for each product.
  + For each product\_id, calculate the average sentiment\_score.
* **Ranking and Filtering:**
  + Rank the products in descending order based on the calculated total revenue.
  + Filter the results to isolate the top 5 products, creating the final dataset for the report.

#### Data Analyst / Business Analyst Responsibilities

The Analyst interprets the results and communicates them effectively to drive business decisions.

* **Interpretation and Reporting:**
  + Take the final aggregated data (Top 5 products with revenue and sentiment) from the Data Scientist.
  + Summarize the findings in a clear, non-technical report.
* **Visualization:**
  + Develop a simple dashboard or chart to visualize the results, such as a bar chart comparing the revenue of the top 5 products, with their average sentiment scores annotated on each bar.

### 3. Flow of Responsibilities and Dependencies

The project's success relies on a clear sequence where each role's output is the input for the next.

1. **Business Analyst → Data Engineer:** The process starts with the Business Analyst defining the requirements. The Data Engineer depends on this to understand what data is needed and how it must be structured.
2. **Data Engineer → Data Scientist:** The Data Scientist depends on the Data Engineer to provide a clean, reliable, and integrated dataset. The analysis cannot begin until the ETL process is complete.
3. **Data Scientist → Business Analyst:** The Analyst depends on the final, analyzed results from the Data Scientist to create the report and visualizations that will be presented to stakeholders.

### GitHub Link

The final report and all associated source code would be stored in a GitHub repository.

* https://github.com/Jagadeeshgowdayt/Data-Engineering-Lifecycle-Lab1.git