**Case Study:** *Exploring the Data Engineering Lifecycle and Stakeholder Roles*

**USN:** 1RVU23CSE300  
**Name:** Nandini P. Dodawad  
**Experiment No.:** 1  
**Date:** 06-08-2025

**Objective:**

To explore the data engineering lifecycle and simulate the roles of key stakeholders (Data Engineer, Data Scientist, Business Analyst) using sample datasets to analyze a business problem and plan a data-driven solution.

**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

**Stage 1: Review Datasets:**  
Reviewed sales\_data\_raw.csv (product\_id, sale\_price, quantity) and customer\_feedback.json (product\_id, sentiment\_score, review\_text) to identify useful fields and patterns.

The company has raw sales data and customer feedback data. Management wants to understand:

* Which are the top 5 products by total revenue?
* How do customers feel about these products?

**Business Question:**  
"What are the top 5 products by revenue, and how does customer sentiment vary for them?"

**Required Data Points:**

* From sales data: product\_id, sale\_price, quantity
* From feedback data: product\_id, sentiment\_score, review\_text

**Requirements Document**

**Business Use Case:**  
To support decision-making on inventory and marketing strategies, the company needs insights into product performance and customer perception.

**Key Metrics to Analyze:**

* Total revenue per product (sale\_price \* quantity)
* Average customer sentiment per product

**Desired Insights:**

* List of top 5 products by revenue
* Average sentiment score and sample feedback for each top product

**Purpose of the Document**

* This requirements document ensured that the **Data Engineer** focused on correct fields and cleaning steps, while the **Data Scientist** applied the right metrics—keeping the solution aligned with business needs.

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

**1. Data Engineer Responsibilities:**

* Ingests raw data from sources like CSV and JSON into tools like Jupyter Notebook or data pipelines.
* Cleans the data:
  + Removes currency symbols (e.g., from sale\_price).
  + Converts string fields to numeric.
  + Handles missing values (e.g., fills missing quantity with defaults).
  + Removes duplicates or corrupted rows.
* Transforms data:
  + Calculates fields like revenue = sale\_price × quantity.

Prepares and stores cleaned data for further analysis by the Data Scientist

**2. Data Scientist Responsibilities:**

* Performs exploratory data analysis (EDA) to identify trends and patterns.
* Calculates key metrics:
  + Total revenue per product.
  + Average sentiment score per product.
* Identifies top 5 products based on revenue.
* Analyzes customer sentiment linked to those products.
* May use visualizations or build simple models to support decision-making.
* Prepares meaningful insights for business teams.

**3. Business Analyst Responsibilities:**

* Reviews output from the Data Scientist to extract key product and sentiment trends.
* Interprets results in business context (e.g., high revenue + low sentiment = risk).
* Highlights performance gaps or customer pain points.
* Designs reports and dashboards to present findings.
* Suggests actionable strategies like:
  + Promoting top-rated products.
  + Fixing issues in poorly-rated high-revenue products.

**Stakeholder Contribution Flow**

* **Data Engineer** → Ensures clean, structured, and usable data.
* **Data Scientist** → Derives insights from data through analysis and modeling.
* **Business Analyst** → Translates those insights into strategic recommendations.

Raw Data

↓ (Cleaned by)

Data Engineer

↓ (Analyzed by)

Data Scientist

↓ (Interpreted by)

Business Analyst

↓ (Acted upon by)

Management / Decision Makers

* If the **Data Engineer** makes mistakes, downstream accuracy suffers.
* If the **Data Scientist** misanalyzes data, insights are misleading.
* If the **Business Analyst** misinterprets results, decisions are flawed.  
  =Clear communication and role coordination are essential for success.

**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.