Case Study Analysis: Data Science Lifecycle

In this case study, you are required to go over the following tasks:

Task 1: Understanding the Business Problem Start by understanding the business needs and objectives:

- Why is the company interested in integrating these data sources?
- What business questions will the data science team aim to answer (e.g., predicting customer churn, optimizing inventory levels)?
- How will data engineering support the data science efforts to generate actionable insights?

Deliverable: Write a brief explanation of the business problem, focusing on how data engineering fits into the broader goal of enabling data science at the company.

Task 2: Designing the Data Pipeline

In this task, you will design a data pipeline that can manage and process data from multiple sources. Consider the following:

- **Data Sources:** Identify the various data sources (transactional, customer, inventory, etc.) and describe how they will be connected to the pipeline.
- **Data Storage:** Plan how and where the data will be stored. Will you use a relational database, NoSQL database, or data lake? Explain the rationale for your choice.
- **Data Integration:** Describe the strategy you would use to integrate the data from different sources into a unified dataset. How will you handle structured and unstructured data?

Deliverable: Draw a diagram of the proposed data pipeline architecture, including data sources, data storage, and how data will flow through the pipeline. Provide a written description explaining your design choices.

Task 3: Data Cleaning and Transformation Strategy

Data from multiple sources often comes in different formats and may have inconsistencies or missing values. In this task, outline the steps you would take to clean and transform the data:

- What data cleaning techniques would you apply (e.g., handling missing values, removing duplicates, normalizing formats)?
- How would you transform the data to ensure consistency across the dataset (e.g., matching field names, standardizing formats)?

Deliverable: Provide a high-level plan for the data cleaning and transformation process. Include the specific steps you would take to ensure data quality and readiness for analysis.

Task 4: Defining the ETL Process Now, focus on the ETL (Extract, Transform, Load) process:

- **Extract:** What tools or technologies would you use to extract data from the various systems (e.g., SQL for relational databases, API calls for third-party data)?
- **Transform:** How would you ensure that the data is transformed to a format suitable for analysis (e.g., data types, aggregations)?
- **Load:** Where will the data be loaded after transformation (e.g., data warehouse, data lake)? What considerations will you make for loading efficiency?

Deliverable: Provide a detailed outline of the ETL process, including the specific tools or frameworks you would use (e.g., Apache Airflow for workflow scheduling, SQL for querying, Pandas for data transformation).