# Data Engineer| 2-4 Years| Mumbai

**Problem Statement**

You are tasked with building a robust and scalable data engineering pipeline that integrates and processes data from multiple sources. The final deliverable should be a standardized dataset that can be deployed in a cloud environment. This assignment will evaluate your ability to design, implement, and deploy a data processing pipeline using industry best practices.

**Required Skills**: Python, SQL, Cloud Architecture, Database, Documentation, Pipelines, Git, Cloud Services

**Assignment Breakdown**

**1. Data Sources:**

You are provided with the following data sources:

**Flat Files (CSV Files)**

* **File Name:** sales\_data.csv
* Transaction\_ID (int)
* Product\_ID (int)
* Quantity (int)
* Price (float)
* Transaction\_Date (date)

**External API (Exchange Rates)**

* **Endpoint**: /api/exchange\_rates
* Currency\_Code (string)
* Exchange\_Rate (float)
* Date (date)

**Internal API (Customer Demographics)**

* **Endpoint**: /api/customer\_data
* Customer\_ID (int)
* Customer\_Name (string)
* Age (int)
* Gender (string)
* Location (string)
* Date\_Joined (date)

**Database**

* **Table Name**: products
* Product\_ID (int) - Primary Key
* Product\_Name (string)
* Category (string)
* Price (float)
* Stock\_Available (int)
* **Table Name**: transactions
* Transaction\_ID (int) - Primary Key
* Customer\_ID (int) - Foreign Key
* Product\_ID (int) - Foreign Key
* Quantity (int)
* Transaction\_Date (date)
* Total\_Amount (float)

**2. Objective**

Your task is to create a data processing pipeline that ingests data from all these sources, standardizes it into a unified format, and performs the necessary data preprocessing steps. The processed data should be suitable for further analysis or downstream applications.

**Steps to Complete the Assignment**

*For Step 1 & Step 2 Please share python scripts wherever needed:*

**Step 1: Ingestion**

Sample data for all the sources (CSV, JSON) is attached with the email (if not please ask the relevant person).

* Write scripts to fetch data from all the provided sources.
* For flat files, use Python to read and parse the CSV files.
* For the APIs, write scripts to extract data, ensuring that proper authentication (token-based authentication) and error handling mechanisms are in place.

You can use Postman (or any other tool) to create mock APIs for testing the internal and external APIs.

* For the database, Import the shared CSV files in a PostgreSQL and use SQL or an ORM to query and retrieve the necessary data.

**Step 2: Standardization**

* Design a data model that standardizes the data from all sources into a consistent format. Consider differences in data types, naming conventions, and structures between the sources.
* Implement the logic to transform and merge data from different sources into this standardized format.
* Share the export of prepared database in .SQL format (or any other relevant format) with all the tables.

*For Step 3, Step 4 & Step 5 Please share the approach documentation and detailed flow like architecture diagrams, Pseudo code, Name of cloud services and their connections:*

**Step 3: Data Preprocessing Pipeline**

* Implement a data preprocessing pipeline that handles missing data, duplicate records, inconsistent data entries, abnormal values (prices less than 0).
* Perform any necessary feature engineering, such as converting categorical variables, normalizing data, or aggregating information.
* Ensure that the pipeline is modular and can be easily extended or modified.

**Step 4: Cloud Architecture**

* Design a cloud-based architecture to deploy your data processing pipeline.
* Also, Add services & architecture to expose a few APIs on top of the prepared database. E,g, **/prices/products/** (returns list of products along with their prices)
* Choose a cloud provider (e.g., AWS is preferred, GCP, Azure) and justify your choice.
* Deploy the pipeline in the cloud, ensuring it is scalable, resilient, and cost-efficient.

**Step 5: Documentation and Presentation**

* Provide detailed documentation that explains your data model, the pipeline architecture, and the steps taken during the implementation.
* Include diagrams or visual representations of your cloud architecture and data flow.
* Prepare a short presentation (5-10 slides) summarizing your approach, challenges faced, and solutions implemented.

**3. Evaluation Criteria**

* **Data Ingestion**: Ability to fetch and handle data from multiple sources efficiently.
* **Standardization**: Quality and consistency of the data model and transformation logic.
* **Pipeline Implementation**: Robustness, modularity, and efficiency of the data preprocessing pipeline.
* **Cloud Deployment**: Suitability, scalability, and resilience of the chosen cloud architecture.
* **Documentation & Presentation**: Clarity, thoroughness, and professionalism.

**4. Submission Guidelines**

* Submit your code in a GitHub repository, with clear instructions on how to run the scripts for steps 1 & 2.
* Include a README.md file that details the steps you followed, the tools used, and any assumptions made.
* Upload your presentation as a PDF in the repository.