**Data Engineer Recruitment - Case Study**

**Income Insurance Limited**

**Objective:**

Company ABC, a dynamic e-commerce platform selling electronics, liquor, and books, seeks to leverage its vast data spread across multiple source systems. The objective is to establish a robust data warehouse and analytical infrastructure to drive data-informed decisions.

**Tasks:**

**1. Data Ingestion & Processing**: Load the data from raw files into relational tables with well-defined schema. Unify the books\_data.csv and books\_ratings.csv into a single flat table. Store all the outputs into relational tables.

**2. Sales Data Mart Creation:** Develop a sales data mart that segments data by product categories (electronics, liquor, books). This data mart should facilitate specific reporting needs such as total sales by region, product category, and more detailed insights.

**3. Automation and Orchestration:** Implement automated workflows using an appropriate orchestration tool (eg: Airflow) to ensure regular updates and maintenance of the data warehouse and data marts without manual intervention.

**Technical Requirements:**

* **Data Processing with PySpark:** Utilize PySpark for all data processing tasks to handle large volumes of data efficiently across distributed systems.
* **Orchestration Frameworks:** Employ robust orchestration frameworks like Airflow or tool of your choice to manage workflow automation seamlessly.
* **Database Management:** Utilize open-source SQL databases for storing all input and output tables, ensuring efficient and ease of integration.
* **Data Mart/Curated Tables:** Sales data mart should be built out of the transaction level information. The sales data mart should provide insights on:
  + Total sales by product categories over different time frames (daily, weekly, monthly, yearly).
  + Total sales by region.
  + Top products by sales volume across various periods and regions.
* **Dashboard Development:** Build a dashboard using tools such as Tableau / Power BI or BI tools of your choice to visualize key business metrics and insights based on the insights to be derived from Data Mart above

**Bonus Points:**

* **Cloud Integration:** Architect the solution using cloud services (AWS/ GCP / Azure) to enhance scalability, performance, and reliability.
* **Data Enrichment for Analytics :** Deriving information from existing fields :
  + Classifying the electronic products into categories like laptops, mobile phones etc. can help business to get a drill down view of insights.
  + Derive state, country information from the address to enable analysis at region level
  + Categories columns imputation using simple NLP techniques for **books** data

**Deliverables:**

* Detailed architecture diagram of the end to end data pipelines.
* Code repository link (Git) with all scripts and implementations.
* Documentation covering the setup, configuration, and usage of the orchestration workflows.
* A demo dashboard providing actionable insights based on the data mart.

**Evaluation Criteria:**

* Technical proficiency in PySpark and SQL.
* Effective use of orchestration tools to create maintainable and scalable workflows.
* Implementation of best coding standards to make sure data pipelines are easily maintainable and scalable as the data and business requirements grow.
* Adherence to best practices in data modelling, ETL processes, and system architecture.

**Dataset:**

You can download the dataset here : [DE\_CaseStudy\_Dataset](https://1drv.ms/f/s!AhsDQYYJJPW7uwlRZA98W7vtPgsc?e=cNMFOD)

**Metadata :** 

This case study is designed to gauge your ability to handle large-scale data infrastructure projects and provide actionable insights through sophisticated data solutions. Good luck!