Assignment 1 – Data Engineering

Name**:** Anshul Bhilare
Enrollment No**:** ADT23SOCB0154

Answer 1: Experiential Learning

Research and identify real-world data sources and integration with tools like Power BI or modern data platforms.

In today's data-driven world, organizations rely on diverse real-world data sources to make informed business decisions. Some key sources include:

- **Retail Domain:** Point of Sale (POS) systems, customer loyalty programs, e-commerce transactions, inventory management systems, social media data, and supply chain systems.
- **Healthcare Domain:** Electronic Health Records (EHR), medical imaging systems, IoT-enabled health devices, insurance claim data, patient feedback systems, and open health datasets (such as WHO, CDC).

Integration with Power BI and Modern Data Platforms:

- Power BI supports direct integration with SQL databases, Excel sheets, cloud data warehouses (Azure Synapse, Snowflake, BigQuery), and APIs.
- In retail, sales transaction data can be ingested from POS systems into Azure Data Lake, transformed using Azure Data Factory, and visualized in Power BI dashboards.
- In healthcare, EHR data can be securely integrated using HL7/FHIR APIs, transformed for compliance, and visualized in dashboards for patient outcome monitoring.
- Modern platforms such as Databricks, AWS Redshift, and Google BigQuery allow scalable ingestion, transformation, and analysis of structured/unstructured data, seamlessly connecting with visualization tools like Power BI.

Thus, real-world data sources, when integrated with modern data platforms and visualization tools, enable organizations to gain actionable insights and drive innovation.

Answer 2: Case Study – Mini Project (Retail Domain)

Problem Statement:
A retail company wants to analyze its sales performance, inventory status, and customer behavior to optimize decision-making.

Data Lifecycle Mapping:

1. **Data Capture:** Sales transactions from POS, customer feedback via surveys, and product inventory records.
2. **Data Storage:** Data stored in SQL Server database and CSV files.
3. **Data Processing:** Using Python (Pandas/NumPy) and Azure Data Factory pipelines for cleaning and transformation.
4. **Data Analysis:** Exploratory Data Analysis (EDA) using Python and aggregated views prepared in SQL.
5. **Data Visualization:** Power BI dashboards to visualize KPIs such as daily sales, top-performing products, inventory shortages, and customer sentiment trends.

Mini Project Deliverables:

- Sales dataset integrated with inventory and customer data.
- ETL scripts for data transformation.
- Power BI dashboard with KPIs and charts.

GitHub Project Link:
https://github.com/anshbytecode/Data_Engineering_anshul

This project demonstrates the complete data lifecycle from capture to visualization, enabling better insights and decision-making for retail businesses.