Name: -Vedant Sasane

Enrollment No.: -ADT23SOCB1284

Assignment 1

# Answer 1: Research and Identification of Real-World Data Sources

Real-world data sources are datasets obtained from practical, real-life scenarios that organizations use to gain insights, make informed decisions, and improve operations. These sources include:  
- Transactional databases (sales, billing, customer records).  
- Open datasets (WHO, Kaggle, Government portals like data.gov).  
- IoT devices and sensor data (wearables, smart meters, medical devices).  
- APIs (social media APIs, financial APIs, weather data).  
- Enterprise systems (ERP, CRM, EHR).  
  
Integration with modern data platforms and BI tools:  
- Power BI: Connects with CSV/Excel files, SQL databases, APIs, and cloud services (Azure, AWS, GCP).  
- Cloud Data Warehouses: Tools like Snowflake, BigQuery, and Azure Synapse allow scalable storage and analytics with seamless integration to Power BI.  
- ETL (Extract, Transform, Load) processes: Used to clean, transform, and load raw data into BI tools.  
  
Conclusion: The ability to capture, integrate, and analyze real-world data with Power BI or modern platforms enables organizations to transform raw data into actionable insights, essential for informed decision-making.

Additional Real-World Data Sources:  
- Social Media Data: Twitter, Facebook, LinkedIn data provide insights into customer sentiment.  
- Web Scraping: E-commerce sites, news portals, and research websites.  
- Healthcare Records: Clinical trials, hospital records, and wearable devices.  
- Financial Market Data: Stock prices, cryptocurrency exchanges, and interest rate data.  
  
Advanced Integration Approaches:  
- Data Lakes: Unstructured and structured data stored together (e.g., Azure Data Lake, AWS S3).  
- Real-Time Data Streaming: Kafka and Azure Event Hub for real-time analytics.  
- Machine Learning Pipelines: Integrating predictive analytics models into BI dashboards.  
- Security & Governance: Role-based access, data masking, and compliance (GDPR, HIPAA).

# Answer 2: Mini Project – Retail Sales Data Lifecycle

Domain Chosen: Retail  
  
Title: Retail Store Sales Analytics  
  
Objective: To design a data lifecycle that captures, processes, and visualizes sales data to provide insights into product performance, revenue trends, and customer preferences.  
  
Steps:  
1. Data Capture: Collect transactional sales data (CSV/Excel file containing product ID, sales date, price, region).  
2. Data Storage: Store the dataset in a relational database or cloud storage.  
3. Data Processing: Use Python (Pandas) to clean and transform the dataset (remove duplicates, handle missing values).  
4. Data Visualization: Import processed data into Power BI and create dashboards showing:  
 - Sales trends over time  
 - Best-performing products  
 - Revenue distribution across regions  
  
Deliverables:  
- Dataset: sales\_data.csv  
- Python ETL Script: etl\_script.py  
- Power BI Report: Retail\_Sales\_Dashboard.pbix  
- Documentation: README.md explaining lifecycle.

Enhancements to the Project:  
- Data Capture Expansion: Include customer demographics, online/offline sales channels.  
- Storage Optimization: Use a star schema in data warehouses for efficient querying.  
- Processing Enhancements: Automate ETL with Apache Airflow or Azure Data Factory.  
- Visualization Enhancements: Use drill-down charts, KPIs, and AI insights in Power BI.  
  
Additional Use Cases:  
- Customer Segmentation: Group customers by purchasing behavior.  
- Inventory Management: Predict stock requirements using past trends.  
- Marketing Insights: Measure campaign success using sales uplift analysis.  
- Forecasting: Apply time series models (ARIMA, Prophet) for future sales prediction.  
  
Conclusion: By expanding the dataset, improving storage and processing, and applying advanced visualization, the project can evolve into a robust business intelligence solution providing actionable insights for retail businesses.