MZUZU UNIVERSITY FACULTY OF SCIENCE, TECHNOLOGY AND INNOVATION
DEPARTMENT OF INFORMATION AND COMMUNICATION TECHNOLOGY

Data Virtualization Assignment 1 Due: 20th June, 2025

Task : Building a Simple Data Warehouse from Multiple Sources

Objective:

Design and implement a basic data warehouse that integrates data from at least three different data sources, transforms the data, and loads it into a centralized PostgreSQL-based data warehouse for querying and reporting.

Instructions:

You have been provided with two files:

1. Data Sources (3 total)

   Use the provided files: profiles.JSONL and sales_data.CSV, and the API is available at the following link: https://dummyjson.com/docs/products

2. Data Extraction

Write Python scripts (using pandas, requests, or any tool) to extract data from the chosen sources.

3. Data Transformation

Perform necessary cleaning and transformation:

- Remove duplicates and nulls
- Format dates and currency
- Standardize naming conventions (e.g., all column names lowercase)

4. Data Loading

Load the cleaned data into a PostgreSQL data warehouse:

- Design fact and dimension tables (e.g., fact_sales, dim_customer, dim_date and any other table that makes sense from the given data)
- Use SQLAlchemy or psycopg2 to load data

5. Reporting

Write SQL queries to generate insights, such as:

- Total sales by region
- Average purchase per customer
- API data correlation with sales (e.g., weather vs. store visits)

6. Documentation

Submit:

- Source code
- ER diagram of the data warehouse
- A short report explaining your ETL process and key findings from your queries