Week 2: Data Ingestion.

Objective: In this lab session, you will get familiar with different data sources and be able to ingest the data using Python, SQL and the command line. We will use Google Colab - Python IDE for most of our tasks.

Go to https://colab.research.google.com/ to start a new Python notebook. You will need to log in to use the Google Colab. You can use your university email or your personal.

To start, make sure you download all the .csv, .xlsx and .json files from Blackboard.

Exercise 1. Ingesting data from flat files (CSV, Excel, TXT, JSON)

```
#import all the useful libraries
import pandas as pd
import requests
import json
```

```
# CSV
csv_url = '/boston_house_prices.csv'
df_csv = pd.read_csv(csv_url, skiprows=1)
print("First Few rows of the CSV Data:")
print(df_csv.head())
```

```
# Excel
excel_url = '/Employee Data.xlsx'
df_excel = pd.read_excel(excel_url)
print("Last Few rows of the Excel Data:")
print(df_excel.tail())
```

```
# TXT
txt_url = 'https://www.gutenberg.org/files/1342/1342-0.txt'
response = requests.get(txt_url)
txt_data = response.text
print("First 200 characters of the TXT Data:")
print(txt_data[:200]) # Print first 200 characters
```

```
#JSON

# Opening JSON file
f = open('/content/sample_users_with_id.json')
# returns JSON object as a dictionary
data = json.load(f)
# Iterating through the json
given_names = [user["given_name"] for user in data if "given_name" in
user]
# Print the names
print(given_names)
# Closing file
f.close()
```

Exercise 2. Ingesting data from a database

For this exercise, we will import .csv files, write them to SQL tables in an SQLite database and ingest them from the database.

i. Load all the .csv files as a pandas dataframe

```
import pandas as pd
import sqlite3
import matplotlib.pyplot as plt
import seaborn as sns
# read csv files
df_BP = pd.read_csv('/content/BUSINESS-PROCESS.csv')
df_TD = pd.read_csv('/content/timedim.csv')
df_PR = pd.read_csv('/content/phonerate.csv')
df_LO = pd.read_csv('/content/location.csv')
```

ii. Create a connection with the sqlite database (Create a new sqlite db if one doesn't exist)

```
# connect to database
conn = sqlite3.connect("DEdb")
cur = conn.cursor()
```

iii. Write the dataframes to sql tables in the database created

```
# load dataframes into database As tables with names BP,DT,PR, LO
df_BP.to_sql("BP", conn)
df_TD.to_sql("DT",conn)
df_PR.to_sql('PR',conn)
df_LO.to_sql('LO',conn)
```

iv. Write a SQL query to ingest from the SQL tables

```
# print all data of the BP tables
BP = pd.read_sql('SELECT * FROM BP', conn)
BP
```

```
# See all data from the DT table
TD=pd.read_sql('SELECT * FROM DT', conn)
TD
```

Exercise 3. Ingest data from the application program interface (API)

For this exercise, we will import and extract weather data from OpenWeatherMap. Kindly sign up for the platform to have an API Key, allowing access to extract data.

i. Ingest the current weather data for London

```
import requests
import json
#API Documentation: https://openweathermap.org/current

currentweather =
requests.get('https://api.openweathermap.org/data/2.5/weather?q=London&
appid=b29954cc60e659dc3ea53752ae96aba4')
print(currentweather.status_code)

print(currentweather.json())

# Save the JSON data to a file
with open('currentweather_data.json', 'w') as file:
    json.dump(data, file, indent=4)
```

ii. Ingest the 5 day / 3 hour forecast data for London

```
import requests
import json
#API Documentation: https://openweathermap.org/forecast5
# Make a GET request to the API
ForecastData =
requests.get('https://api.openweathermap.org/data/2.5/forecast?q=london
&appid=b29954cc60e659dc3ea53752ae96aba4')
print(ForecastData.status_code)
print(ForecastData.json())

# Save the JSON data to a file
with open('ForecastData.json', 'w') as file:
    json.dump(data, file, indent=4)
```

Extra Activities: Use the Command Line to move around files

i. Create a folder named DE_WK1.

ii. Move all the downloaded files (.csv, .xlsx, .json and .db) to the newly created folder.

iii. Check your newly created folder to confirm if the files were moved.