Practical Exercise: Integrating Weather Data Into a Sales Dataset Using APIs and MongoDB

Problem Statement

You work for a retail company that suspects weather patterns (e.g., temperature, rainfall) may influence sales. Your task is to build an integration pipeline that adds weather data to the sales data, allowing the company to analyze sales patterns against weather conditions. The sales data is stored in a CSV file with information about transaction dates, product codes, and sales amounts. You will use the OpenWeatherMap API to fetch historical weather data for the location of the retail stores, integrate it with the sales data, and store the result in MongoDB.

Data Sources

- Sales Data (CSV)

  You are given a CSV file named sales_data.csv found on the following link:https://github.com/DrManalJalloul/Introduction-to-Data-Engineering/blob/main/sales_data.csv containing the following columns:

  - date: The date of the transaction.
  - product_id: The ID of the product sold.
  - sales_amount: The total amount of sales for that product.
  - store_location: The city where the store is located.

- Weather Data (OpenWeatherMap API)

  You will use the OpenWeatherMap API to retrieve historical weather data for each city where the retail store is located. For each date, you will fetch the temperature, humidity, and weather conditions. Example API request (replace YOUR_API_KEY with your actual API key):

  https://api.openweathermap.org/data/2.5/weather?q=NewYork&appid=YOUR_API_KEY

Step-by-Step Instructions

Step 1: Extract Sales Data from CSV

The first step is to extract the sales data from the CSV file. This will serve as the main dataset to which we will add weather data.

Step 2: Fetch Weather Data from the API

Next, use the OpenWeatherMap API to fetch weather data for each store location on the corresponding transaction date.

API Setup:

- Sign up for an OpenWeatherMap API key at OpenWeatherMap.
- Use the requests library in Python to pull weather data.

Check the following sample python code:

```
import requests

# Function to fetch weather data for a given city
def fetch_weather_data(city, date, api_key):
    base_url = f"http://api.openweathermap.org/data/2.5/weather?q={city}&appid={api_key}
    response = requests.get(base_url)
    data = response.json()

    # Extract temperature, humidity, and weather description
    temperature = data['main']['temp'] - 273.15  # Convert from Kelvin to Celsius
    humidity = data['main']['humidity']
    weather_description = data['weather'][0]['description']

    return temperature, humidity, weather_description

# Example usage:
api_key = 'YOUR_API_KEY'
temp, humidity, description = fetch_weather_data('New York', '2023-09-01', api_key)
print(f"Temp: {temp}, Humidity: {humidity}, Weather: {description}")
```

Step 3: Combine Weather Data with Sales Data

Now, you'll need to combine the sales data with the weather data. Loop through each row of the sales dataset, retrieve the weather data for the location and date, and add it as new columns in the dataset. Note that you will need to create the new columns: Temperature, humidity, and weather_description in your sales_data dataframe.

Step 4: Load the Integrated Data into MongoDB

Finally, you will load the integrated sales and weather data into MongoDB for future analysis.