ETL Exercise

The ETL Process will be used to carry out the following exercise in Data Analysis.

We have a set of customer sales data that we want to extract information from.

Initial setup

1. Setup mysql database with docker. To do this, run `docker-compose up` command from the handouts directory.
2. Create 'test' database in your MySQL.
3. Create virtual environment with python, install any required packages such as pymysql or pandas. Activate the virtual environment.

Task

Write a python script or a jupyter notebook that executes the below steps.

Extract

1. Extract all the data from the sales_data.csv. The columns for the csv are `customer_id`, `purchase_date`, `purchase_amount` and `product_id`.

Transform

1. Clean that data (minimum requirement is to remove any rows that contain null cells).
2. Filter data for the period 1 December 2020 - 5 December 2020
3. Calculate each customer's total spend
4. Calculate each customer's average spend
5. Calculate how many times each customer has purchased a specific item

Load

1. Load the results into a database. You can choose how to structure the table(s) you load the data into but an example was given to you in sql-utils.py file. If using pandas, consider using .to_sql dataframe method.

Analyse

1. What does the data in the tables tell you about the different customers purchasing habits?