TWO PARTS

PART 1

Prepare to present you knowledge about data-architecture: e.g star schema and OLAP cubes.

PART 2

Create a simple data pipeline, you can use tool of your choice (Any programing language). Provide small documentation how to run your solution and put your code on github. Steps the program should perform:

1. Get data:
    - Combine the primary datasets (task_data_1.csv, task_data_2.csv) into a one common dataset. The number of datasets is unknown in advance, the program must be able to process N datasets without additional modifications. Datasets are all located in the same folder ../task_data
    - By using policy_id column join POLICY_DATA table to the dataset, which was made in the first step.
2. Data processing. Requirements:
    - Replace missing values of "car_brand" by mode (value that appears most often in a set of data values)
    - Create a new column - "car_age". Use "car_registration_year" and the date of today for calculation. After creating this variable.
    - Replace missing value of "car_eng_pow" by "-1". Then group values into "0-100", "100-250", "250+" groups.
    - Regroup "marital_status" in a binary way: 1 - Single, 2 - Married. Blank fields should be grouped into a Single group.
    - Replace missing values of "claim_amount" with "0".
3. Final data saving
    - Save final dataset into a file in results folder (../results)
    -