

Task 2: Data Munging

Here is the background information on your task

Your team has been tasked with collecting metrics on a plethora of disparate shipping data. This task comes straight from the top, so it would be wise to give it your all. The data is contained in a number of different spreadsheets, each with its own competing schema. In order to interrogate the data, all of it has to be in the same place and in the same format. Currently, the shipping data exists in several places in several formats and is therefore impossible to query. To combine the spreadsheets, you need to write a python script to read through every row, extract the pertinent data, figure out how to combine it, munge it into the right format, and upload it to the database. Plenty of steps, but the resulting data will be much easier to query. Once the database contains all the data, you can pass it off to the analysis team to extract all the relevant metrics. Good luck!

Here is your task

Part 1: Get the data

First, you need to get your hands on the relevant data. The shipping department has been kind enough to provide you with a repository containing all of their spreadsheets, as well as a copy of the sqlite database. First, fork and clone the repository at: <https://github.com/theforage/forage-walmart-task-4>

Part 2: Populate the database

Your task is to insert all of the data contained in the provided spreadsheets into the SQLite database. You will write a Python script which:

- Reads each row from the spreadsheets.
- Extracts the relevant data.
- Munges it into a format that fits the database schema.
- Inserts the data into the database.

Spreadsheet 0 is self-contained and can simply be inserted into the database, but spreadsheets 1 and 2 are dependent on one another. Spreadsheet 1 contains a single product per row, you will need to combine each row based on its shipping identifier, determine the quantity of goods in the shipment, and add a new row to the database for each product in the shipment. The origin and destination for each shipment in spreadsheet 1 are contained in spreadsheet 2. You may assume that all the given data is valid - product names are always spelled the same way, quantities are positive, etc.

When you're finished, convert the python script you used to populate the database into a PDF and submit it below. You'll see an example answer on the next step, but we encourage you to give it a go first!

Question Solution Explanation:

Step 1: Fork and Clone the Repository

- Fork the repository from the provided GitHub link:
<https://github.com/theforage/forage-walmart-task-4>
- Clone the forked repository to your local machine using Git

Command:

```
git clone https://github.com/your-username/forage-walmart-task-4.git
```

```
cd forage-walmart-task-4
```

Step 2: Install Required Libraries

Ensure you have the necessary Python libraries installed. You'll likely need `pandas` for handling spreadsheets and `sqlite3` for interacting with the SQLite database.

Command:

```
pip install pandas sqlite3
```

Step 3: Examine the Provided Files

Take a look at the spreadsheets and the SQLite database schema to understand what kind of data you're working with and how it's structured

Step 4: Writing the Python Script

I have placed my Python code for this data munging task in the **scripts** folder within the Task 4 directory as **data_munging.py** file.

Step 5: Running the Script

- Place your script in the same directory as the spreadsheets and run it.
- This script reads each spreadsheet, processes the data according to the instructions, and inserts it into the SQLite database.

Step 6: Convert the Script to a PDF

Once you've tested and confirmed that the script works, convert the Python script to a PDF.