Data Engineering

The goal of the exercise is to give you a real example of the type of data engineering that we work on. Our basic pipeline involves collecting, cleaning, and organizing data from different sources. These data pieces then need to be integrated into a single database model.

Included are the following files:

- volume16.pdf
  - This is a pdf of public transaction data from the Malaysian Land Department
  - The objective is to ingest data that comes in this format
- volume16-raw.csv
  - This is a sample of what the extracted version should look like
  - The objective is to create this file
-

Your objective is to create a data pipeline (using python3 and a database – postgresql is recommended) that accepts as an input the pdf file of transactions (similar to the volume16.pdf file) and ingest it into a data model that allows users to query for:

- the min, max, average/median number of transactions
  - by geographic area
  - by property type
  - by land use
  - by tenure
  - by date
- the min, max, average/median transaction price
  - by geographic area
  - by property type
  - by land use
  - by tenure
  - by date
-

User Acceptance Criteria

1. Design and document a data pipeline that ingest new listings data
   a. There should be an ER diagram of the tables involved
   b. There should be a diagram of the data flow and process flow
2. The python program should have a command line interface and accept a new pdf file
3. The ingestion pipeline should result in the transaction data loaded into the appropriate tables

4. Any “bad” records should be identified for future review
5. There should be no “duplicate” records in the various tables (document how you identify duplicates)
6. To review the assignment, we will be running a different set of pdf transaction records using the cli
    a. Provide instructions on how to setup your pipeline
    b. Assume that we have a Postgres DB available

Notes

1. In the real-world, data sets are not complete – thus you will have to decide what to do with records that do not match or are missing
2. The data sets in this assignment are not “cleaned” there will be dirty records. The system should identify them
3. If you have questions or are not certain about something – then write it down. We are looking to see how you think about the pipeline
4. If you can dockerize the solution that would be preferred