. 1. You are required to make a user-agent that will crawl the WWW (your familiar domain) to produce dataset of a particular website.
   - the web site can be as simple as a list of webpages and what other pages they link to
   - the output does not need to be in XHTML (or HTML) form
     a multi-stage approach (e.g. produce the xhtml or html in csv format)

(10 marks)

For this task, I have chosen to crawl properties listing in Kuala Lumpur on mudah.my. 10 pages of properties listings were crawled and subsequently 40 urls collected from each page were obtained to crawl individual ads. A total of 400 ads were acquired with various attributes such as price, area, title info, property type, facilities and so on.

Please refer to jupyter notebook for implementation:
Q1_FinalExam_WQD180113_ChoongEnJun.ipynb

2. Draw snowflake schema diagram for the above dataset. Justify your attributes to be selected in the respective dimensions.

(10 marks)

The dataset crawled from Q1 is not complex, and I can only justify one additional dimension to make it a snowflake schema from what otherwise will be a star schema.

The additional dimension, “property_type” is extended from the first parent dimension called “category”. It is a logical extension of the dimension because “property type” is a subset of “category” as is “bungalow” is a subset of “houses”

The snowflake schema is established by having this additional dimension, and it reduces data redundancy thus saving disk space. By having this additional dimension, we effectively reduced one column form the fact table.

Please refer to jupyter notebook and sqlite3 file for implementation: Q2_FinalExam_WQD180113_ChoongEnJun.ipynb

Attachment: Snowflake schema visualized using DBeaver on the sqlite3 file