Capstone Data Storage Solution

Following is a summary of the data storage of this project

Type	Form	Format
Raw data	Data files	CSV
Processing data	Data files	Parquet
Processed Data	Data files	Parquet
Data for query	RDBMS table	

1. Raw data
   When the data files are downloaded from the source, they will be stored at a separate location, such as a folder called 'raw' or similar. They will remain in original format and be stored as csv files.

2. Processing data

   During the cleansing/transformation process, the data will be stored in a special location , could be a folder called 'processing' or similar. The data will saved as parquet files

   Why we choose Parquet, is because it is one of the columnar formats which have storage and performance benefits. The values are clustered by column so the compression is more efficient (to shrink the storage footprint), and a query engine can push down column projections (to reduce read I/O from network and disk by skipping unwanted columns), otherwise known as column pruning. Parquet also stores the file schema in the file metadata so when it is read, the readers can have data and schema at the same time

3. Processed data
   Once the transformation ends, the data will be stored in a specific location, waiting to be loaded into data warehouse. The data format will also be parquet

4. Data for query
   Eventually, the data will be loaded into data warehouse, which is built on a relational database, for use. They will be in form of dimension tables or fact tables. We will create indexes which will