Capstone Data Storage Solution

Here is the summary of the data storage of this project

Type	Form	Format
Raw data	Data files	CSV
Processed data	Data files	Parquet
Integrated data	Data files	Parquet
Published data	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 be saved in the original format, csv.

2. Processed data

   During the transformation process, we will cleanse data and perform necessary transformations, such as removing bad records, changing data format, removing columns that are not needed, etc . At the end of the process, , the data will be stored in a location which can be considered as the processed layer. The data will be 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. Integrated data

   As we are targeting building up a data warehouse, we will further process the data by conducting data warehouse modeling and create new datasets .These datasets will be equivalent to the dimension tables and fact tables that are in a star schema. They should be ready to be loaded to a RDBMS based data warehouse system or serve as data files of the external tables if that is the choice. These files will be stored in format of ‘parquet’

4. Published data

   Eventually, the data will be exposed to the end users for query, analysis or other purpose.

   There are two ways to do this:

   1. To load the data into a RDBMS based data warehouse system. The users can directly run SQL query or do analysis on these data, using various types of tools

   2. Create external tables in the data warehouse system (such as Azure Synapse, Databricks, AWS Redshift), which will be built on to these data files,

      This way, the data will remain in the data lake on the cloud storage, such as S3 or Azure blob, which are of low cost.. And the users can also access the data by running SQL.