ETL Project Report

Team: Taylor Walraven and Daniel Viassolo

- **Extract**

  Original data sources:
  - US Traffic Fatalities for 2015 and 2016 by State and City (referenced by Geographic Locator Code)
    - https://www.kaggle.com/usdot/nhtsa-traffic-fatalities
    - This is a BigQuery Dataset, no files to download, but can be queried using the BigQuery API
  - US Geographic Locator Codes - .xlsx
    - https://www.gsa.gov/reference/geographic-locator-codes/glcs-for-the-us-and-us-territories
    - Data format: Excel
  - Historical Weather Data (including 2015-16) for 30 cities in US
    - https://www.kaggle.com/selfishgene/historical-hourly-weather-data/data
    - CSV

- **Transform**

  At a high level these were the steps involved in the Transform phase:
  1. Extract the 30 city names from the Historical Weather Data source
  2. Extract city geographic locator codes from the Geographic Locator Codes source for all 30 cities
  3. Use Kaggle BigQuery to extract traffic fatalities that
     a. occurred in those cities
     b. occurred during the years for which we have weather data (2015-16)

  All the details involved in the Transformation phase can be found in the Jupyter Notebook *extraction_and_transformation.ipynb* (included in the Github repo). This notebook is used to generate the 3 csv files:
  - fatalities.csv
  - weather.csv
  - location.csv

  to be later loaded in a database in the Load phase.

- **Load**

  A PostgresSQL database was defined and the data from the above 3 CSV files was loaded into 3 tables with same names – fatalities, weather, location.

The Entity Relationship Diagram (ERD) is included on the Github repo (PNG file). The “schema” file is also included in the Github repo.

The main reason to choose a Relational (or SQL) Database over a non-SQL database (e.g., MongoDB) is that the dataset in question is well-structure. SQL database have in general better performance than their non-SQL counterparts.

Possible future applications of this database include:

- Study correlation between accident severity (number of fatalities) and weather conditions
- Study correlation between accident severity (number of fatalities) and time of the day (that could be representative of traffic volumes)
- Study correlation between accident severity (number of fatalities) and geographic locations (that could be representative of driving habits)