|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | | |

Ken Ho 9/23/2020

ETL Project Description

This project involves creating a database showing the flight delay and flight distance covered by flights originating for all cities in the USA.

The source of the data is

https://github.com/PacktPublishing/Learning-PySpark/tree/master/Chapter02/flight-data

ETL Process

There are two source files as follows:

1. departuredelays.csv -> it has flight delays and distance for multiple origins and destinations in the USA

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **date** | **delay** | **distance** | **origin** | **destination** |
| 1011245 | 6 | 602 | ABE | ATL |

1. airport-codes-na.csv -> it has mapping between Airport Codes (IATA) and cities

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **City** | **State** |  | **Country** | **IATA** |
| Abbotsford | BC |  | Canada | YXX |

These files are available in *Data* folder attached.  These source files were first converted from txt to csv. The data was loaded into MongoDB.

We joined both datasets to get City, State, Country, Origin and Destination Airports. Then datasets were prepared for output Table 1 and Table 2 as follows:

Table 1: sum of flight delay for each destination city and in decreasing order of delay

Table 2: total distance covered by flights originating for all cities in the USA and average delay for each city

To run the mode, we needed pymongo module. It can be installed using the command below:

conda install -c anaconda pymongo

The output tables are in MongoDB. The Jupyter notebook is also attached.