CIS 9440 – Data Warehousing for Analytics
Final Project Milestone 3
April 24th, 2021
Group #2
Student: David Freitag

ETL Process Summary

For this milestone of the project, I wrote a Python script to perform extraction of all data from various sources, cleaning and transformation of that data, and loading of that data into BigQuery. Here are the steps involved in this process:

1. Create a new BigQuery project (Chicago Crime Project) in Google Cloud Platform
2. Create a new dataset for the Chicago Crime Project (chicago_crime_project_data)
3. Create a service account and download the access key
4. Write etl_functions.py and use ETL.ipynb to test these functions
5. Once all ETL functions are working correctly, export ETL.ipynb to ETL.py
6. Run ETL.py to perform all ETL operations

Here is a breakdown of the files used in this process, which are included in the project directory:

Filename	Description
chicago-crime-project-311420-0a513b4ce130.json	Service account access key
etl_functions.py	ETL functions called in ETL.py
ETL.ipynb	Jupyter notebook used for developing and testing ETL functions
ETL.py	Main ETL script used to perform all ETL operations
script_output.txt	Output from the terminal produced when ETL.py was run

The operations in ETL.py are as follows:

Extract:

- Extract Chicago crime data from the BigQuery public dataset for the years 2011-2019
- Extract Chicago Public Schools graduation data from the Chicago Public Schools website (note: only years 2011-2019 are included in the data, which is why I only selected those years from the other data sources, despite data for other years being available for extraction)
- Extract Chicago local area unemployment rate from the Bureau of Labor Statistics (BLS) using a library that utilizes the BLS data API for years 2011-2019
- Extract Chicago temperature and precipitation for years 2011-2019

Profile:

- Run the profiling function on all four datasets to observe a summary of the data, columns, memory usage, etc.

Clean

- Run a cleaning function on each dataset to drop rows with nulls and remove duplicates

Transform:

- Create the date dimension using a SQL query to generate a date array
- Create the crime code dimension from the Chicago crime data by removing duplicate values and isolating necessary columns
- Create the location dimension from the Chicago crime data by removing duplicate values and isolating necessary columns
- Create the crime_incident fact by merging the dimensions with the Chicago crime data and dropping the unnecessary columns
- Create the chicago_unemployment fact by transforming the rate for each month into a daily rate that is consistent for every day of the month it applies to (i.e. every day in Januay 2011 has the same unemployment rate, which equals the rate for that month as a whole)
- Create the graduation_rate fact by transforming the rate for each year into a daily rate that is consistent for every day of the year it applies to (similar in process to the chicago_unemployment fact)
- Create the weather fact by adjusting the date format into the same format as the date dimension's date_id column

Load:

- Load all dimensions into BigQuery
- Load all facts into BigQuery

While the above description provides a summary, the clearest way to observe how the ETL pipeline actually functions is to look at the code itself in conjunction with the script_output.txt file, which lists out the operations being performed.

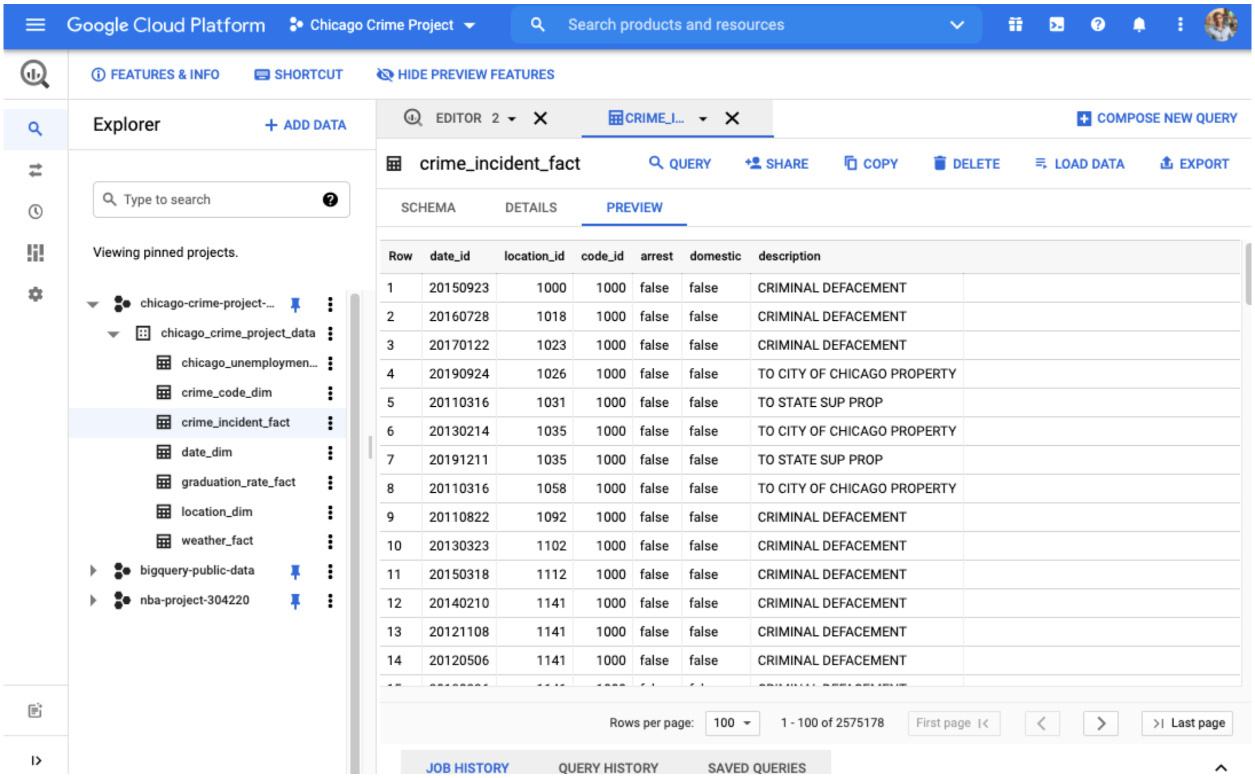

Figure 1.1: *A preview of the crime_incident_fact table in Google BigQuery with the rest of the tables visible.*