ETL Project Report

10/1/20

EXTRACT

I utilized 2 data sources:

OpenWeather (History Bulk)
https://openweathermap.org/history-bulk
Downloaded CSV
Query: Folsom Lake Coordinates (38.7206, -121.1339), Weather Hourly, 8/31/17 – 8/31/20

Department of Water Resources California Data Exchange Center
https://cdec.water.ca.gov/dynamicapp/wsSensorData
Downloaded CSV
Query: FOL, RESERVOIR ELEVATION (hourly), 8/31/17 – 8/31/20

TRANSFORM

For the weather data, I imported the CSV into a Jupyter Notebook and converted it to a dataframe. The timestamp was in the Unix Epoch format. I converted this column into the 'datetime' format. I also renamed the column in preparation for joining it. I then created a streamlined version of the dataframe that included the columns I wanted in the new database: date/time, temperature, humidity, wind speed, weather type, and cloudiness.

For the lake elevation data, I imported the CSV into a Jupyter Notebook and converted in to a dataframe. The timestamp column that was available appeared to be in the right format, but upon closer inspection I realized that would need to be converted from a string to the 'datetime' format. I renamed 2 of the columns in preparation for joining the table. I then created a streamlined version of the dataframe that only included the date/time and elevation level.

I joined the two cleaned dataframes utilizing merge on the 'date_time' field. It was clear that there was discrepancy between the count of rows between the two dataframes, so I ran a check to see if there were any empty values present after the merge. I identified that there was some elevation data missing from the CDEC dataframe for a handful of hours throughout various dates. To avoid uploading empty data to the final database, I dropped any rows that had empty elevations.

LOAD

I decided on MongoDB for the database for this project. Given the tool's flexibility and performance advantages, I believe it was the right choice for the 26k+ records that needed to be loaded. I uploaded the final merged dataframe into a single Collection and printed the results.