Lab Report

Title: Lab 1
Notice: Dr. Bryan Runck
Author: Yaxuan Zhang
Date: 10/10/2023

Project Repository: https://github.com/YaxuanSeanZhang/MGIS_ARCGIS/tree/main/GIS%205571/Lab1
Google Drive Link:
Time Spent: 8 hrs

Abstract

stands for Extract, Transform, Load, and it is a crucial process in data integration and data warehousing. ETL is used to gather, process, and transfer data from various sources to a data warehouse or other target systems for analysis and reporting. In this lab, we will build ETL pipeline to extract data from different APIs for different data types. By comparing different ETL workflows, we will gain a deeper understanding of ETL process.

Problem Statement

In this lab, we will go through the ETL process for different data types, e.g., shp, geojson, and csv. Through practicing decomposing interfaces for spatial web API's into informal conceptual models, we can compare contract different web API's using informal conceptual models and custom-built ETL routines. We will build an ETL pipeline with ArcPro Jupyter Notebook and integrate two datasets via spatial join.

Table 1. Main Steps

#	Requirement	Defined As	(Spatial) Data	Attribute Data	Dataset	Preparation
1	Extract Data from API	Raw input dataset pulling from different API (i.e., Geospatial Commons, Google Places, NDAWN)	Polygon, Point	Various	Minnesota Geospatial Commons, Google Places, NDAWN	Define API, and pull data
2	Transform coordinate	WGS 1984	Polygon, Point	Various	Minnesota Geospatial Commons, Google Places	Convert to shapefile and then transform the coordinate
3	Spatial Join	Arcpy spatial join	Point joined Polygon	Various	Minnesota Geospatial Commons, Google Places	1 point joins 1 polygon

Input Data

We will use data from three sources: Minnesota Geospatial Commons, Google Places, NDAWN. Data from Minnesota Geospatial Commons is shapefile data. Data from Google Places are formatted as geojson. Data from NDAWN are csv tables. We need to customize the API to define the range of the data, and then pull data via customized API.

Table 2. Required Dataset

#	Title	Purpose in Analysis	Link to Source
1	Census2020CTUs	Boundary Data	Minnesota Geospatial Commons
2	Restaurant Nearby	Nearby restaurant data	Google Places
3	Average Weather	Daily average weather of one station	NDAWN

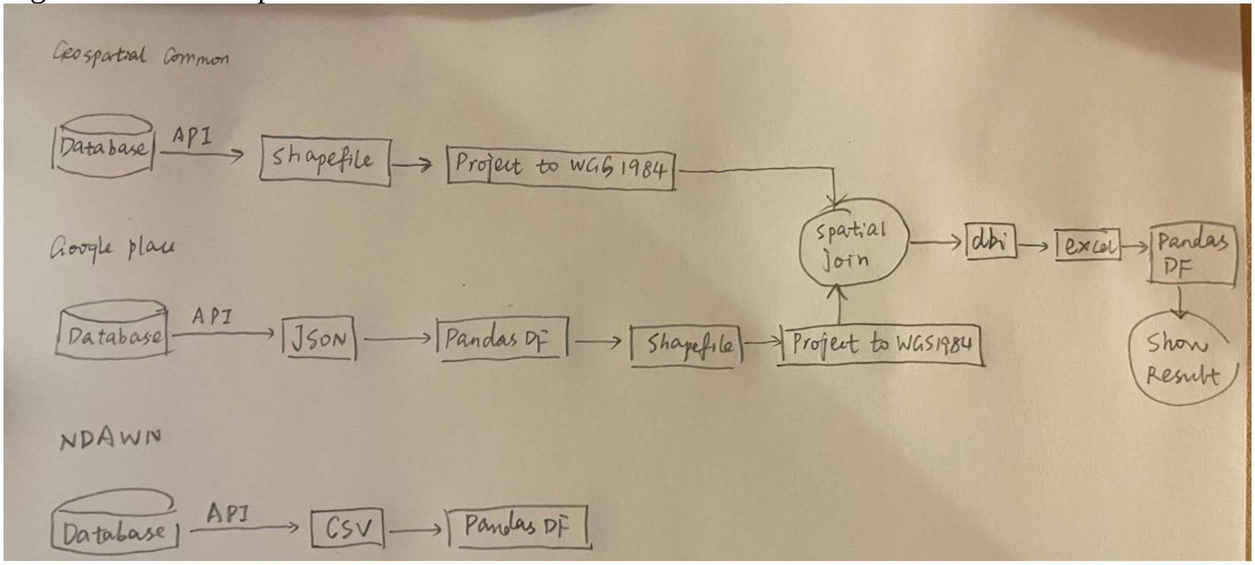
Methods

Minnesota Geospatial Commons (shp): Basically, you right click the download button and copy the address, that would be the api. For shapefile data, you need to further unzip the zip file.

Google Places (json): You need to sign up for the google cloud account and get the api key. Then you create the api by defining location, radius, keyword, etc. Next, you need to convert json into data frame and shapefile if you need to perform spatial analysis.

NDAWN (csv): You need to create the api by defining station, variable, type, begin date and end date.

Figure 1 ETL and Spatial Join Workflow



Results

The results are shown in Jupyter Notebook, e.g., the head of shapefile table (Minnesota Geospatial Commons), the first item of geojson (Google Places), the head of transformed data table from geojson (Google Places), the head of the spatial joined table (Minnesota Geospatial Commons & Google Places), and the head of csv table (NDAWN).

Results Verification

I ran the whole workflow and there was no error. Also, I visualized the extracted data and make sure they are under the same coordinate.

Discussion and Conclusion

I mainly learned how to modify the api to define what range of data you want to extract. By using the requests package, we can directly extract via the defined api.

I also learned how to use JSON data and how to convert it into data frame and shapefile in order to further perform spatial operations.

In summary, I had a good understanding of how to build an ETL pipeline.

Self-score

Category	Description	Points Possible	Score
Structural Elements	All elements of a lab report are included **(2 points each)**: Title, Notice: Dr. Bryan Runck, Author, Project Repository, Date, Abstract, Problem Statement, Input Data w/ tables, Methods w/ Data, Flow Diagrams, Results, Results Verification, Discussion and Conclusion, References in common format, Self-score	28	28
Clarity of Content	Each element above is executed at a professional level so that someone can understand the goal, data, methods, results, and their validity and implications in a 5 minute reading at a cursory-level, and in a 30 minute meeting at a deep level **(12 points)**. There is a clear connection from data to results to discussion and conclusion **(12 points)**.	24	24
Reproducibility	Results are completely reproducible by someone with basic GIS training. There is no ambiguity in data flow or rationale for data operations. Every step is documented and justified.	28	28
Verification	Results are correct in that they have been verified in comparison to some standard. The standard is clearly stated **(10 points)**, the method of comparison is clearly stated **(5 points)**, and the result of verification is clearly stated **(5 points)**.	20	20
		100	100