# Lab Report

Title: GIS5571 Lab1
Notice: Dr. Bryan Runck
Author: Tzu Yu Ma
Date: October 10, 2024

**Project Repository:** https://github.com/TzuYuMa/GIS5571/tree/main/Lab1
**Time Spent:** 14 hours


## Abstract

This Lab is to build a pipeline to use APIs for data retrieval from Google Places, Minnesota Geospatial Commons, and NDAWN, performing coordinate reference system transformation, and conducting special data joins. Understanding the significance of ETL (Extract, Transform, Load) and constructing the appropriate ETL pipeline is crucial. ETL plays a pivotal role in modern data management and analysis. It is the process of transforming raw data into a format suitable for analysis, reporting, and decision support. An ETL pipeline is the cornerstone of the project's success, facilitating the extraction, transformation, and loading of data from multiple sources into valuable information assets, laying a solid foundation for further analysis and decision-making. Therefore, prioritizing and constructing the appropriate ETL pipeline in the project is of paramount importance.


## Problem Statement

Identifying pertinent data and streamlining its organization can pose a challenge. Data sourced from different APIs often exhibit diverse structures and may prove excessively large to manage efficiently. In this project, I will use Google Places, Minnesota Geospatial Commons, and NDAWN these three APIs to request the data. The primary aim is to leverage the ETL process to tackle these challenges and render the data suitable for utilization within GIS applications.

*Table 1. Requirement Information of the Project*

| # | Requirement | Defined As | (Spatial) Data | Attribute Data | Dataset | Preparation |
|---|---|---|---|---|---|---|
| 1 | Raw data | Request from Google Places, Minnesota Geospatial Commons APIs | geometries | Depends on the data requested | Google Places, Minnesota Geospatial Commons | Request the ideal results from APIs |
| 2 | Transform the projection | Transform the raw data to the same projection (In this project I use WGS 84) | geometries | Depends on the data requested | Google Places, Minnesota Geospatial Commons | The raw data is in JSON/CSV format, need to convert to GDF first |

| | | | | | | |
|---|---|---|---|---|---|---|
| 3 | Spatial join the datasets and display | Use GDF to inner join the data. And display on Folium map. | geometries | Depends on the data requested | Google Places, Minnesota Geospatial Commons | Make a Folium map |
| 4 | save into a geodatabase | GDF to the feature which can use in GIS application | geometries | Depends on the data requested | Google Places, Minnesota Geospatial Commons | Convert GDF to feature layer |

**Input Data**

Retrieve data from three different APIs:

Google Places API: Obtain information on nearby restaurants within a specific location and radius (in this project the location set to Minneapolis and a radius of 1000 meters). The data is retrieved in JSON format.

Minnesota Geospatial Commons API: Retrieve data related to the Functional Class Roads. The data is retrieved in JSON format.

NDAWN API: Retrieve weather data for various states and specific dates. The weather data is obtained in CSV format. This data will be collected but not used in the subsequent processes.

*Table 2. Data Sources*

| # | Title | Purpose in Analysis | Link to Source |
|---|---|---|---|
| 1 | Google Places API | collecting location-based data from Google Map | Google Places |
| 2 | Minnesota Geospatial Commons API | Functional Class Roads contains major roads and highways in the Twin Cities seven county metropolitan area | Minnesota Geospatial Commons |
| 3 | NDAWN API | Weather data for the US | NDAWN |

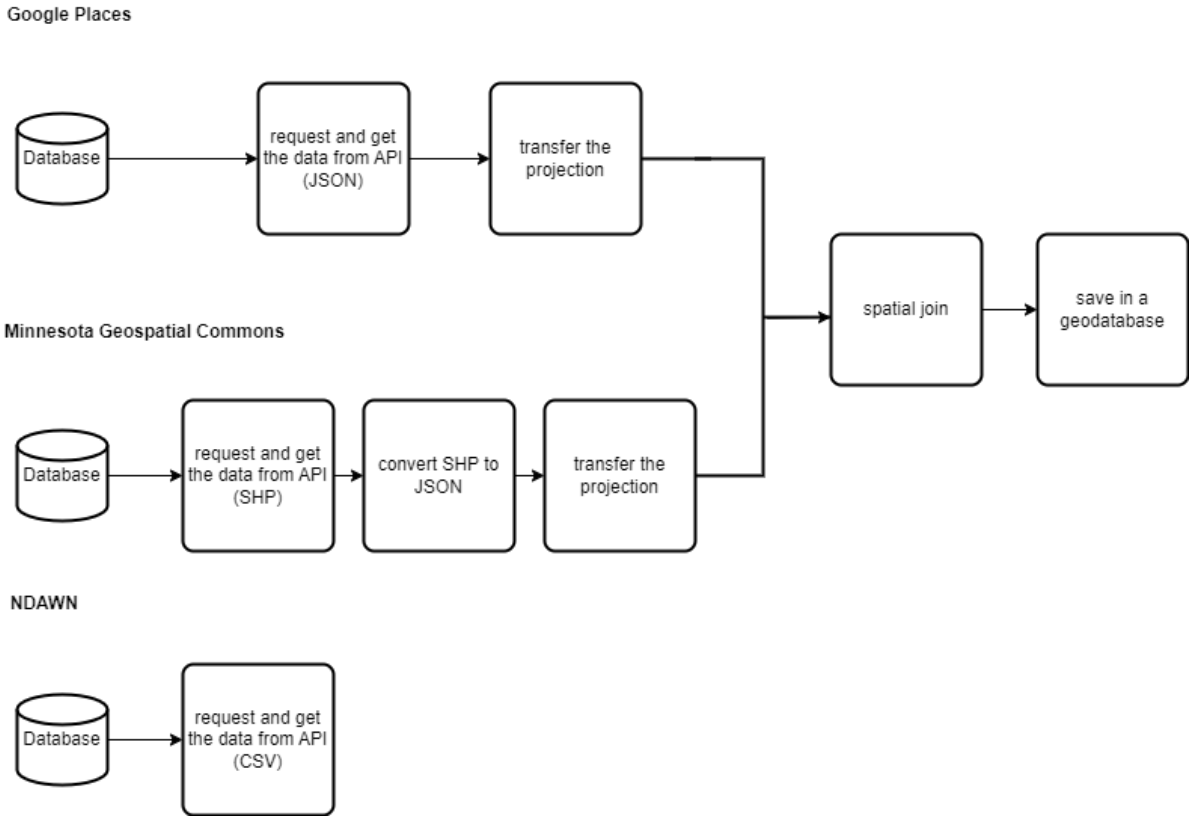
**Methods**

Google Places: Locate the API, request the data, and save it as a JSON file. In this project, the goal is to find nearby restaurants within a 1000-meter radius.

Minnesota Geospatial Commons: Fetch the data via the data's URL in JSON format. In this project, the data pertains to the exciting functional class roads contains major roads and highways in the Twin Cities seven county metropolitan area.

To ensure uniformity and ease of organization, convert both datasets into Geo Data Frames (GDF) and subsequently transform the coordinate projection to WGS 84. After standardizing the coordinate system, employ Geopandas to perform a spatial join between these two datasets.

Ultimately, utilize Folium to create a map and visualize the results.




**Results**

The result is a map that displays nearby restaurants within a 1000-meter radius of the campus, along with county boundaries. The map's central location was set at coordinates [44.9, -93.2].

**Results Verification**

I display the result by using Folium to create an map, providing a visual representation of the data. This not only confirms that I have acquired the desired dataset but also verifies the successful completion of the join operation. Additionally, the map enhances the interpretability of the combined data, making it more accessible and informative for further analysis or presentation purposes.

**Discussion and Conclusion**

I encountered obstacles right from the beginning. I wasn't familiar with APIs and how to request data from them. However, as I delved into researching the methods, I gained a better understanding of how to access the data I needed. I also realized that the most effective way to learn something new is to dive in and do it.

Building the ETL pipeline was a novel experience for me, and I initially had to research what ETL entails. This research helped me grasp the essence of the task at hand. In the final step, I faced some challenges while saving the spatially joined data into the geodatabase. It seemed like I encountered obstacles at every step of the process. However, I also noticed that the initial ETL pipeline was the most challenging, the second one was manageable, and the third one became smoother and more efficient as I gained experience.

## References

Burlingame, E. (2023, September 7). How to use an API: Just the basics. TechnologyAdvice. https://technologyadvice.com/blog/information-technology/how-to-use-an-api/

What is ETL (Extract, transform, load)? Definition, process, and tools. (n.d.). Talend - A Leader in Data Integration & Data Integrity. https://www.talend.com/resources/what-is-etl/

## 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 | **23** |
| **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 | **27** |
| **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 | **18** |
| | | 100 | **96** |