Database Optimization Process and Data Visualization for Public Sector

Haoyan Li, MPA candidate, Sol Price School of Public Policy, University of Southern California
Southern California Association of Governments Department of Research and Analysis,
Randall Lewis Data Science Fellowship

Background

Goods movement is essential to support the economy and quality of life in the SCAG region. The regional goods movement system is a multimodal, coordinated network that includes deep-water marine ports, international border crossings, Class I rail lines, interstate highways, state routes and local connector roads, air cargo facilities, intermodal facilities, and distribution and warehousing clusters.

International Trade data on US Census Bureau's website is an important tool for Goods Movement team better understanding goods movement in SCAG area and doing comparison with other regions across the county. However, the process of acquiring data from the website and doing analysis was extremely inefficient. Goods Movement team desired a database optimization process which makes data retrieval and data analysis be achieved in a more efficient way.

By implementing the data management optimization process, the time for handling data of a certain 2-digit good of one year changed from three weeks to one and a half hour. The efficiency has been largely enhanced. Besides, the data sets were changed into a better structure, which is more user-friendly.

ETL Process

ETL is short for extract, transform, load, the three database functions that are combined into one tool to pull data out of one database and place it into another database. In this database optimization process, there are four steps instead: downloading, converting, cleaning and importing. We need to write codes for extracting data from US Census Bureau's API and cleaning the datasets before importing to MySQL database.

Recommendations

- Popularize database knowledge among public sectors

When doing data analysis, many employees in public sector merely rely on Excel and are not familiar with database or Structured Query Language. Database has many advantages over Excel: the capacity is way higher, manipulation is more diversified, data is organized in a better structure…In order to enhance the efficiency of data-supported decision making, it is necessary to increase staff's knowledge of SQL and database.

- Do regular database maintaining

The construction of database cannot be accomplished overnight. Every month and every year, new data should be imported to the database to guarantee real-time of the information. Maintenance personnel should regularly check the integrity of data.

- Introduce data visualization to data analysis

In public sector, data visualization can also be a critical tool to demonstrate facts and an important propeller to make right policies. Tableau is a powerful Business Intelligence & Data Visualization tool produced by Tableau Software Company. Tableau can be connected to almost any database, people drag and drop to create visualizations, and share with a click. Besides the common plots, Tableau has a mapping functionality, and is able to plot latitude and longitude coordinates and connect to spatial files like Esri Shapefile, KML, and GeoJSON to display custom geography, which is particularly helpful for people who do urban planning.