ETL Project

By

Daniel Ghaffarian

Maria Hudson

Philip Nunoo

Date Submitted: July 20, 2019

# Introduction:

Our team was approached by a client who is evaluating the North American retail markets health and running an analysis for the sales and employment over the last five years. The client has asked our team to extract and transform their data from three different government databases, and migrate it on their database for their analysis.

# Executive Summary:

For this project our team extracted the employment data, sales data, and the North American Industry Classification System (NAICS) codes for each retail category from multiple different files within three different government databases, transformed the data by cleaning and eliminated unwanted narratives, unwanted footnotes, and irrelevant data, joining the different tables using the NAICS codes, and loading the data in the clients SQL database.

1. **Evaluation:**

The data integration for our project took place in the following steps:

* 1. **Finding the Data (Extract)**

In order to find our data we used three different government databases as our sources. Here is the sources we used for our data:

1. **Sales Data for Retail and Food Services (Codes 44-45):** The sales data was extracted for years 2018-2014. The data for each year was extracted from its respective worksheet within the same CSV file. In total 5 different worksheets were used.
   * Source: [United States Census Bureau](https://www.census.gov/retail/index.html)
   * Data: [Sales Data for Retail and Food Services](https://www.census.gov/retail/mrts/www/mrtssales92-present.xls)
   * Data Format: xls/csv
2. **NACIS code and title associated with each code:** The data for the NAICS codes were organized in a table within this website. We used Pandas and Data Scraping technique to extract the data from this source.
   * Source: [naics.com](https://www.naics.com/six-digit-naics/?code=44-45)
3. **Employment Statistics in the Retail Trade Sector:** The employment data was extracted for years 2009-2019. The data for each year was extracted from its respective CSV file CSV file. In total 5 different CSV files were used.
   * Source: [Bureau of Labor Statistics](https://www.bls.gov/iag/tgs/iag44-45.htm)
   * Data: 14 files were used from this source. Please refer to Data\_Input folder on GitHub.
   1. **Data Cleanup and Analysis (Transform)**

After the data was extracted from the mentioned sources, the data had to be cleaned and structured into the desired format. The following transformations were completed in order to transform the data in a desired format:

1. **Cleaning:** Cleaning of the data was completed by eliminating the unwanted narratives, unwanted footnotes, and irrelevant data.

Additionally, there were some combined trade sectors that could not be separated or merged with other tables (e.g. 442, 443 keys were combined into one trade sector).

This could cause an issue when merging the data and therefore was removed from the data.

1. **Transform:** The extracted data for this project was from three different sources and in different formats. The sales data was extracted in separate tables by year, whereas, the employment data was extracted in separate tables by Trade Sector code. As a result the differences in the data formats, we had to create a universal format where the data is organized by NAICS codes, year, and month. Transforming the data in said fashion gives the users the ability to merge the data by NAICS code, month, or year.
   1. **Database Load**