Data Engineering Capstone Project

- US I94 Immigration Analytics

Author: Fan Yuan

Created: 04/25/2021

# Overview

This project attempts to build an ETL pipeline hosted on S3 with the use of Spark so as to set up efficient analytics framework with I94 immigration, global land temperatures and US demographics data. Within the project, we will load data from S3, process the data into analytics tables (fact and dimension tables which will act as the fundamental of further business cases analysis) using Spark based on EDA (exploratory data analysis) result, and load them back into S3. The Spark process will be deployed on a cluster using AWS. The project breaks down into steps as following:

# Scope the Project and Gather Data

* Project scope

This project is going to utilize I94 immigration, global land temperatures and US demographics data to build up analytics database. By combining the datasets appropriately, analysis can be done on the dataset to explore the pattern and insight of US immigration along the time.

The first step will be loading the data set with Spark, then appropriate processing and cleansing will be made based on EDA (exploratory data analysis) on I94 immigration, demographics and global land temperatures dataset. Then, dimension and fact tables will be created accordingly so that further advanced analysis can be done on the database.

* Data to be used

The project will be working with four datasets that resides in S3. The main dataset will include data on immigration to the United States, and supplementary datasets will include data on airport codes, U.S. city demographics, and temperature data.

* I94 Immigration Data: This data comes from the US National Tourism and Trade Office. The data contains international visitor arrival statistics by world regions and select countries (including top 20), type of visa, mode of transportation, age groups, states visited (first intended address only), and the top ports of entry (for select countries).
* World Temperature Data: This dataset came from Kaggle. The data contains the information of global average temperature, average temperature uncertainty by country and city.
* U.S. City Demographic Data: This data comes from OpenSoft. This dataset contains information about the demographics of all US cities and census-designated places with a population greater or equal to 65,000.
* Airport Code Data: This is a simple table of airport codes and corresponding cities. The table also contains the information of airport type, elevation type, gps\_code and coordinates.
* End use cases

The followings are the potential cases which can be analyzed using this data model:

How international traveler distributes within months over the whole year?

How international visitor arrival location expand?

How people choose which port and which travel mode to enter the US?

How does age and visa status of international visitor look?

Is there any relationship between where international visitors come from and the land temperature?

* Technology and tools

The project will store the data on Amazon S3 and use Apache Spark to read in source data from staging tables, extract necessary columns needed for analysis and populate the fact and dimension tables. Then will use Spark to write the data back to S3 if needed. For the data modeling part, the project will use the dimensional model which will make it easy for business users to work with the data and also improve analytical queries performance. So, in this case, we will use Star Schema which fits OLAP (online analytical processing) very well.

* Why S3 and Spark

When dealing with the large dataset in this project, with the combining of both batch and streaming capabilities, Spark can support the use case very well where the data need to be stored and analyzed in real-time. It will have more flexibility when more type and volume of data sources need to be added. Therefore, storing the data on S3 will eliminate need to invest in costly hardware and scale up with full flexibility when needed. And speaking of Parquet files, the columnar format that being used will be a good option to store big data set and for analytics purpures as well.

And Spark can efficiently read data from S3 and process the data with full sets of data analytics and machine learning libraries. Especially when dealing with large dataset, Spark has more capacity to handle the performance and efficiency.

# Explore and Assess the Data

Explore the data to identify data quality issues, like missing values, duplicate data, etc. And clean the data appropriately based on EDA results.

* I94 Immigration Data

By looking at the labels mapping of I94 immigration dataset, we can select some candidate columns which will be used in later data modeling. Since we are only looking at 2016 I94 immigration dataset, there’s no need to keep year column, but will keep the month column so that we might be able to get some insight related to month. I94cit, i94res and Port columns can be kept for exploring the location distribution, mode, birth and visa will be helpful information for analyzing demographic insight. Visapost column can also be useful if we look at where most of international visitor come from, and last will include female and airline columns at first round to see if we can get any valuable information from them.

After checking the missing value of candidate columns, it showed that occup and visapost columns have over 50% missing data which makes the column kind of useless in the later analysis, therefore drop those two columns also. Another basic check here is removing duplicates in dataset if exists. Also, we will drop rows which have missing value in all columns.

* World Temperature Data

First, I checked the range of the date column within the dataset. It turned out that the data only includes the temperature data from 1743 to 2013. Since the I94 immigration data only has 2016 data, and the temperature has the data till 2013, in that case, we will only look at the 2013 temperature data. And the I94 immigration data only have the information at country level for which country the visitor is from and at most state level for which state the visitor enter the US, in the following steps, we will use this data set at country level.

There’s very little amount of missing data in Average Temperature and Average Temperature Uncertainty columns (4%) and since we will only use temperature dataset at country level, we can just fill the missing data with the group average at country, dt level. After filling the NAs with group mean, we still have some missing data which means it doesn’t have records even by group level, then we’ll just remove those records (2%).

* U.S. City Demographic Data

Looking at US city demographic data, first check the overall missing data and it turned out there’s a vert few missing data (0.5% of the whole dataset at most), so here decide to just remove the rows with missing data. And also remove duplicate rows if exists.

* Airport Code Data

Since after initial assessment, it looks like the airport code dataset is difficult to join the other three datasets especially the I94 immigration data (which doesn’t have city level information). There’s no good option for common column to join the dataset for further analysis, therefore, here decide to not be using this dataset in the data modeling.

# Define the Data Model

In this project, we will use Star Schema to convert the data which is a model designed to support OLAP (online analytics processing) at its best. We will have Fact tables which will record business events (like an order, a phone call, or a book review, etc.) recorded in quantifiable metrics (like quantity of an item, duration of a call, or a book rating, etc.). And Dimension tables will record the context of the business event (e.g., who, what, where, why, etc.) with attributes value like location of a store where the item was purchased, or name of the customer who made the call, etc.

* Conceptual data model

The data model will have one fact table that record the key information, and several dimension tables which contains the details. The country dim table has contry mapping and average temperature information. This can support the analysis on how the temperature changing or pattern are related to the immigration trend. The US demographics dimension table grabs information from demographics dataset and connect to the fact table with state level mapping. This will help the analysis on relationship between the immigration pattern and US demographic data. The visa dimension table includes the visa categories information from immigration dataset and can be linked back using visa type key.

Diagram

Description automatically generated

* Steps to map out the data into the model

The steps to pipeline the data into the chosen data model are as following. First, load in the dataset for processing and cleaning, after getting clean immigration data frame, we can extract visa and arrive date dimension table. After loading the global temperature data, create country dimension table and immigration fact table. Then, load in and clean the demographic data then extract and create the demographic dimension table.

# Run ETL to Model the Data

* Create the data pipelines and the data model

Create the arrive date dimension using the arrdate column in immigration dataset and add the other levels of date in the meanwhile. After that, extract the country name and mapping from i94 res file, aggregate the temperature dataset to country level, so that we can get the country dimension table. Then use visatype column in immigration data frame to create visa dimension table.

* Data dictionary

Table

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, application, email

Description automatically generated

* Data quality checks

After loading all fact and dimension tables, check the count of each table to ensure all of the tables are loaded as expected.

* How often the data should be updated and why

The I94 immigration data used in this project is updated monthly, therefore it would be a good choice to update the data model designed in this project monthly as well.

# More application scenarios

* The data was increased by 100x.

Since Spark is designed for handling big data set, the increased data set won't be a big issue for Spark. But it might be an option to change some setting when setup the clusters like node numbers, computer power, etc.

* The data populates a dashboard that must be updated on a daily basis by 7am every day.

We can utilize the Apache Airflow to schedule the pipeline running so that we can get time on time every day.

* The database needed to be accessed by 100+ people.

We can move the database to cloud like using Redshift so that we can support more access better.