

ETL Project



May 16, 2020

By

Bhavna Patadia & Quecis Joshua

# Summary

The scope of this project was to utilize ETL capability to extract information on a specific topic, either via web scraping or importing raw csv data, transforming into consumable components, and loading it into a database so analysis on the topic can be conducted. The topic for our data integration project (ETL) was data pertaining to COVID-19 beginning from May 01 to May 07, 2020. The desire was to get data from multiple sources to see relationship of COVID-19 cases (confirmed cases or deaths) against the state population.

We pulled data from three different sources. The first two sources came from USAFacts.org site (see Sources section). All relevant county population data as well as confirmed cases and death due to COVID-19 was available on the site in CSV format. However, the information within these data files was missing State’s name so a third source was used to get state related information. The data files contained many fields which were not relevant to the scope of the report hence those columns were dropped during the transformation phase. Once all relevant data was finalized, it was then uploaded to the *covid\_data* database using pandas functions.

There were three specific set of information that was analyzed from the data collected

1. Entire dataset of COVID-19 related data – state name, State abbreviation, county name, county population, data collection date, confirmed cases for that day, confirmed deaths for the day [all data extracted is between 5/1/2020 – 5/7/2020]\*
2. Aggregated data at state level including – state name, state abbreviation, aggregated county population, aggregated confirmed cases within all counties in a state, aggregated deaths within all counties in a state.
3. Understanding percent of state population confirmed with COVID-19 showing following data – state name, state abbreviation, aggregate population, aggregated confirmed cases, and calculated percentage of population infected with COVID-19.

# Sources of data

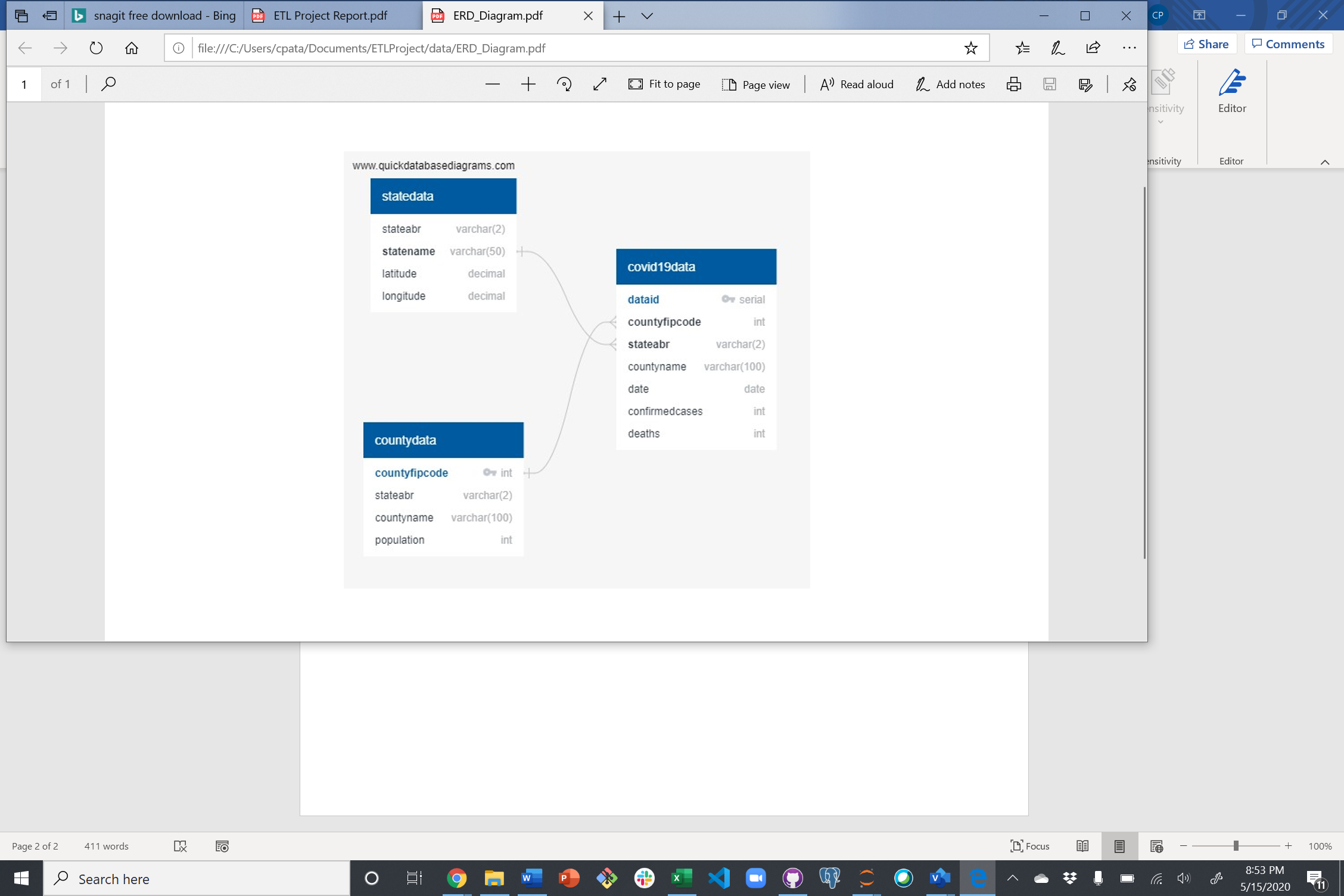
1. USAFacts.org COVID-19: <https://console.cloud.google.com/bigquery?project=planar-catbird-272517&folder=&organizationId=&p=bigquery-public-data&d=covid19_usafacts&page=dataset>

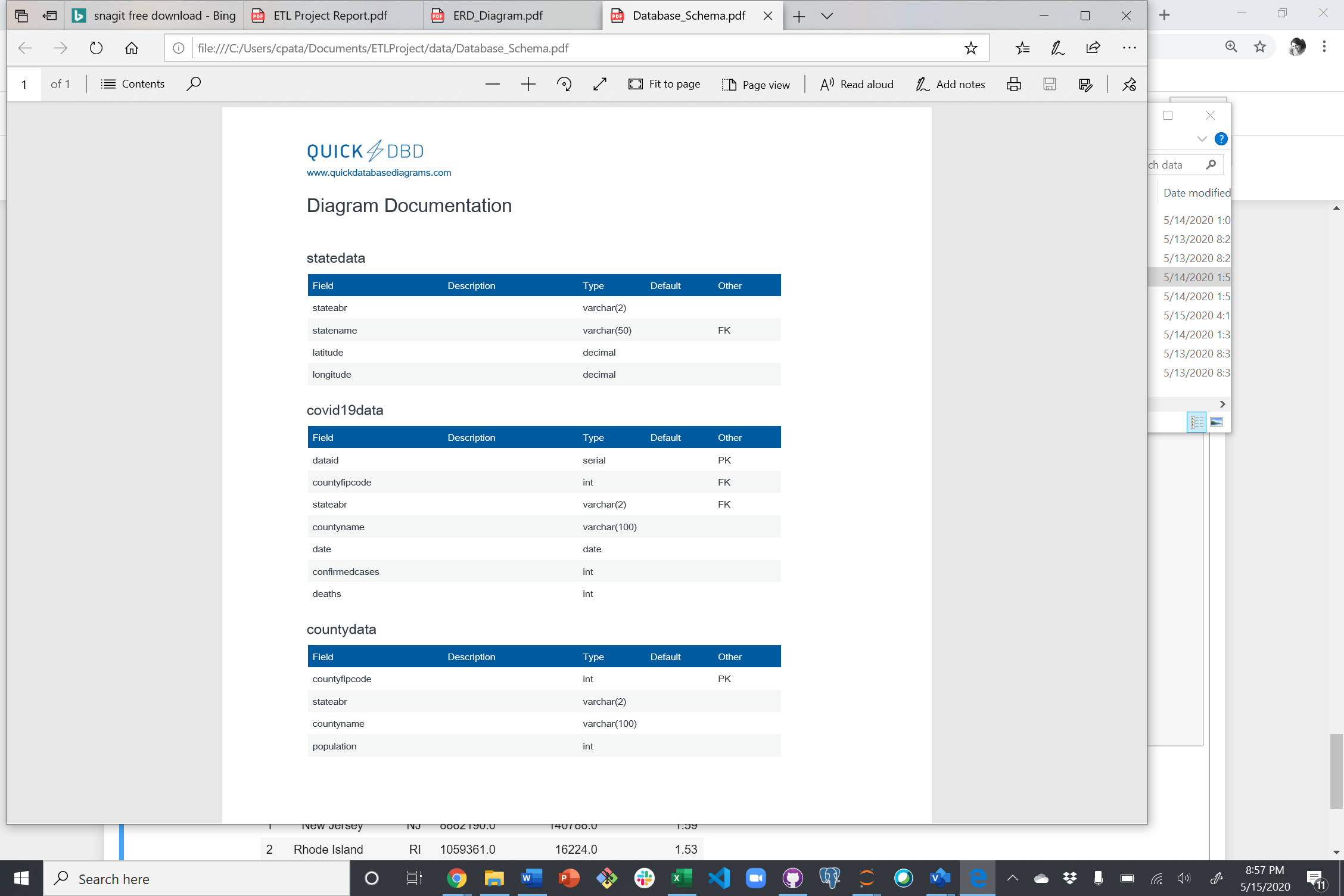
A screenshot of a social media post

Description automatically generated

1. USAFacts.org Population: <https://usafactsstatic.blob.core.windows.net/public/data/covid-19/covid_county_population_usafacts.csv>
2. Google State: <https://developers.google.com/public-data/docs/canonical/states_csv>

# Database

Database schema was identified based on the information that was needed to meet the scope of the project. Hence many columns within the csv files, gathered from source 1 and source 2, were dropped. The relationship between source 1 and source 2 was based on the unique key countyfipscode. The relationship between source 1 & source 2 to source 3 was on state abbreviation. This linked all three sources to make a cohesive set of data as identified in the ERD diagram.

Using the ERD as template, the below database schema was created. Using pandas to\_sql functions, database was populated with all the data.