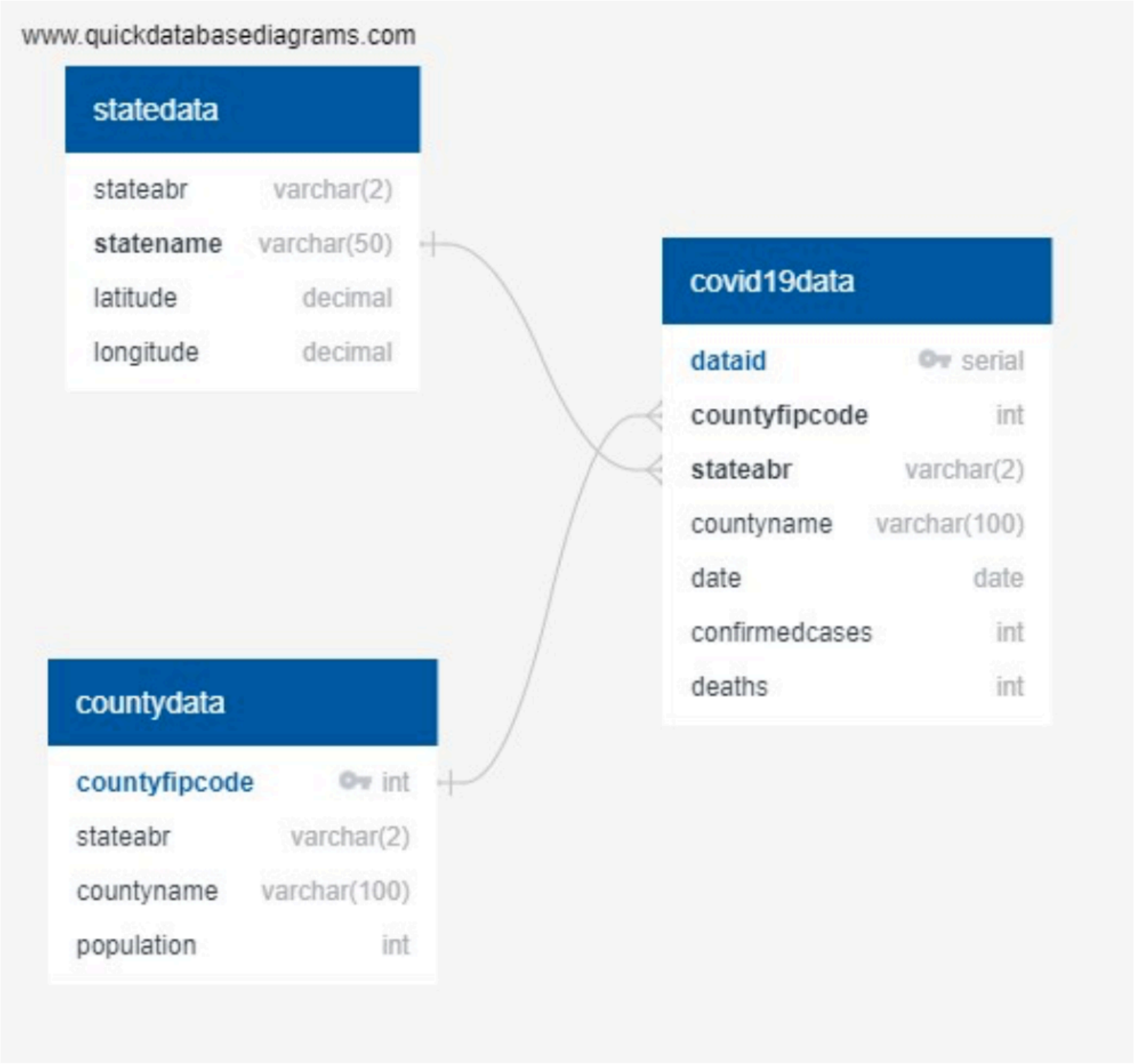
ETL Project Report
Bhavna Patadia
Quecis Joshua
05/16/2020

The Topic for our data integration project (ETL) was COVID-19. We pulled Data from three different sources The first two sources came from https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/ This is where all the data about COVID-19 was stored here we found two CSV file that gave us information on the confirmed cases in each county and state. These CSV files also had the population of every state. We used web scraping to collect the state name and abbreviation of each state this data was found on googles public docs here. https://developers.google.com/public-data/docs/canonical/states_csv We then gathered the data into a Jupyter notebook file and use Pandas to clean and transform the data. Some columns had to be renamed to match the database schemas we created with quick database diagrams, And the tables e created in Postgres. We then used sqlaclhemy to store this information in our database and tables into Postgres where the data was loaded.



www.quickdatabasediagrams.com

statedata	
stateabr	varchar(2)
statename	varchar(50)
latitude	decimal
longitude	decimal

covid19data	
dataid	serial
countyfipcode	int
stateabr	varchar(2)
countyname	varchar(100)
date	date
confirmedcases	int
deaths	int

countydata	
countyfipcode	int
stateabr	varchar(2)
countyname	varchar(100)
population	int

QUICK DBD

www.quickdatabasediagrams.com

Diagram Documentation

statedata

Field	Description	Type	Default	Other
stateabr		varchar(2)		
statename		varchar(50)		FK
latitude		decimal		
longitude		decimal		

covid19data

Field	Description	Type	Default	Other
dataid		serial		PK
countyfipcode		int		FK
stateabr		varchar(2)		FK
countyname		varchar(100)		
date		date		
confirmedcases		int		
deaths		int		

countydata

Field	Description	Type	Default	Other
countyfipcode		int		PK
stateabr		varchar(2)		
countyname		varchar(100)		
population		int		