**Database Structure creation for COVID19\_DATA\_ANALYSIS**

Initial stage:

1. Create new tablespace for creation of database
2. Create new database
3. Create new schema
4. Create table for loading raw data from covid.csv

Pandas ETL stage:

1. Cleanup the loaded dataset for
   1. Uniform date format mm/dd/yyyy instead of dd/mm/yyyy
   2. Cleaning up duplicate rows for ID column
   3. Replacing rows with dates ‘9999-99-99’ to ‘9999-12-31’ later to be changed as 0 during ML transformation.
   4. Any other cleanup changes which shall be encountered further.

**Overview of steps for sourcing data into database/python for ML modelling and Dashboards:-**

1. Load the Raw dataset file into a postgres table
2. Cleanup the data using Pandas
3. Load the clean data back to postgres so that the before and after data is saved in a table, which one can easily query.
4. Create separate tables for dashboard and ML modelling. Good practice is to create separate schema for ETL, dashboards, ML modelling. However, since our project scope is small, we may not require to create those.

**Logical Data model**

Source Data Table:

1. covid\_dataset – Raw data from source file

Metadata Tables:

1. catalogs – Description of codes in column values
2. column\_description – Column description

Staging Tables:

1. master\_covid\_stg – Clean covid data

Master Tables:

1. master\_covid – Clean covid data with primary key and constraints

Child Tables for ML modeling:

1. covid\_patient\_deaths
2. covid\_patient\_infected
3. covid\_patient\_infected\_icu
4. covid\_patient\_infected\_intubed

Child Tables for Dashboard:

1. covid\_preexisting\_conditions
2. covid\_age