Capstone Project
Case Study
COVID-19 Data Processing and Analysis

COVID-19 Data Analysis
Case Study

Business Case Scenario

- Large amount of data is available on COVID-19 for public access from government organizations and other sources like large news agencies in several countries.
- The data includes incidence of the pandemic such as confirmed cases, casualties for example in the US at the state level, county and at region level.
- Now data on vaccination allocation and distribution also is available from these sources which is continuously getting updated.
- Once the data is cleaned and augmented with a few additional fields, several insights can be obtained from the data for any month.

Data Dictionary

- Center for Disease Control (CDC) a US government organization , makes available data on the number of COVID-19 cases, fatalities data-wise on each state on a daily basis.
- It also publishes data on allocation of vaccines from the 3 authorized pharmaceutical companies – Pfizer, Moderna and Jenssen on a weekly basis giving the number of vaccinations allocated and distributed to each state.

COVID-19 Data Analysis
Case Study

Data Dictionary - *COVID-19 Cases and Deaths by State over Time*

- For the case study we have considered the data ***COVID-19 Cases and Deaths by State over Time***.
- The data from the last week of Jan-2020 till the second week of May-2021 is taken up for the data processing and analysis exercises.
- Following are the fields provided in the dataset.
- The data is updated on a daily basis.

submission_date	state	tot_cases	conf_cases	prob_cases	new_case	pnew_case	tot_death

conf_death	prob_death	new_death	pnew_death	created_at	consent_cases	consent_ deaths

COVID-19 Data Analysis
Case Study

Data Dictionary - *COVID-19 Vaccine Distribution Allocations by Jurisdiction*

- The other major data set used is ***COVID-19 Vaccine Distribution Allocations by Jurisdiction***
- US Food and Drug Administration has authorized 3 brands of vaccines.
  - Pfizer
  - Moderna and
  - Jenssen
- Data on allocation and distribution of these vaccines is available from the last week of Decembet-2020 till second week of May-2021 and it is updated on a weekly basis.
- The data dictionary of this data set is as given below.

Jurisdiction	Week of Allocations	1st Dose Allocations	2nd Dose Allocations

COVID-19 Data Analysis
Case Study

Data Dictionary - *Census Data*

- Vaccine allocation and distribution is done based on the population data.
- Population data is available from the US Census organization, with the base data of 2010 and the estimated population of 2019 is taken as the basis by CDC.
- The following fields are available in the dataset from which the required fields are extracted for processing.

COUNTY	STNAME	CTYNAME	CENSUS2010POP	ESTIMATESBASE2010	POPESTIMATE2010	POPESTIMATE2011	… … …	POPESTIMATE2019

COVID-19 Data Analysis
Case Study

Problem Statement

- The first step of processing the data is to clean up the records based on certain criteria.
- Next the data needs to be augmented by adding a few fields to facilitate better data analysis.
- Once the above steps are completed the specified reports need to be generated from the data.
- These steps are explained in detail in the approach provided below.

Approach

- PySpark and Hive are to be used for the project.
- The data can be loaded into a PySpark SQL DataFrames creating as many DataFrames as required by the available data.
- From the census data only the required columns need to be taken and loaded into a look up table to get the population of any state based on the state code.
- A delimited file of the names of states in the US and their standard code also needs to be used as it gives a standardize way to refer to the states in all the datasets.

COVID-19 Data Analysis
Case Study

- For data cleansing we need to check and remove the records where
  - The sum of confirmed cases and probable cases is not equal to the total cases.
  - Similarly records when the sum of confirmed deaths and probable deaths is not equal to the total deaths the records are rejected.
  - The dates given in the first data set are not in the default format of PySpark SQL/Hive. So we need to format the dates to make use of the timestamp/date calculations.
- Next the data needs to be augmented by adding a few derived fields as explained below.
  - Take the census data and extract the estimated population of 2019.
  - Look up the standard state names and code table and get the state-wise census into a DataFrame and/or a Hive table.
  - Once we are equipped with the above processed data we can write the data into Hive tables in Parquet format.

COVID-19 Data Analysis
Case Study

- From this data the following listing or reports need to be generated and displayed.
  1. Compute and list the ***Positive Case Rate per 100000*** population for each state. This is calculated as:

     (Total positive cases in the state / Total estimated Population of the state)*100000
  2. Get the number of identified positive cases in each state in the last 7 days of the reporting period.
  3. Compute and list the ***Deaths Rate per 100000 due*** to COVID-19 population for each state. This is calculated as:

     (Total deaths due to COVID-19 in the state / Total estimated Population of the state)*100000
  4. Get the number of deaths due to COVID-19 as identified in each state in the last 7 days of the reporting period.

COVID-19 Data Analysis
Case Study

- Vaccine Allocation & Distribution
  5. State-wise breakup of allocation of Pfizer vaccine
  6. State-wise breakup of allocation of Moderna vaccine
  7. State-wise breakup of allocation of Jenssen vaccine
  8. Total number of vaccines allocated – State-wise breakup
  9. Ratio of the population covered with vaccinations in each state based on the allocation and population figures from the census data.
  10. Proportion of population that is so far not yet covered i.e. without access to vaccines as on date

COVID-19 Data Analysis
Case Study

- Load the data into PySpark data frames including the look up tables
- To address each of the report requirements use
  - PySpark SQL functions or
  - Write SQL queries using SparkSession
- Store all reports resulting in multiple rows in Hive tables in the default Parquet format
- Show the Hive tables and the table descriptions once the reports are saved