COVID-19 Vaccination Progress Analysis using Google Dataflow

Project Group 6

Vamshi Krushna -015351310

Shwetarani -015277964

Samhitha Upadhyaya-015276092

Introduction

- Countries are suffering from COVID-19 pandemic, they are trying to bounce back by finding a cure in the form of vaccines.
- Countries have been successful in finding different vaccines for this disease. Currently, they are in the process of vaccinating the people with the goal of reaching herd immunity soon.
- As per our dataset there are total 194 countries vaccinating their population with 34 vaccine combinations.
- The project aims to convey the analysis of different ongoing vaccination programs ,by building a data analytics pipeline using Google Dataflow..
- That will provide analysis like the most widely used vaccines and most vaccinating countries...

Dataset Description

- The dataset used is “COVID 19 World Vaccination Progress” for which the data is collected from the “Our World in Data” repository (https://ourworldindata.org/covid-vaccinations).
- The dataset contains Country-by-country data on global COVID-19 vaccinations.
- It also contains daily information regarding vaccinations worldwide from 12th December 2020 to till today and it keeps updating.

The dataset contains the following information:

Country : country for which the vaccination information is provided

Country ISO Code : ISO code for the country

Date : date for the vaccination data entry

Description Cont’d..

Total number of vaccinations : This is the absolute number of total immunizations in the country

Total number of people vaccinated : Distinct count of people who received at least one dose of vaccine.

Total number of people fully vaccinated: This is the number of people that received the entire set of immunization according to the immunization scheme (typically 2).

Daily_vaccinations_raw: Daily change in the total number of doses administered.

Daily_vaccinations: New doses administered per day (7-day smoothed).

Total vaccinations per hundred: Ratio between Total number of vaccination and total population up to the date in the country.

Description Cont’d

People vaccinated per hundred:Ratio between population immunized and total population up to the date in the country.

People fully vaccinated per hundred:Ratio between population fully immunized and total population up to the date in the country.

Daily vaccinations per million:Ratio between vaccination number and total population for the current date in the country.

Vaccines:Different types of vaccines used in the country.

Source name:Source of the information (national authority, international organization, local organization etc.)

Source website:Website of the source of information.

GCP Services used

Data Ingestion

- Google Cloud Storage
- Google Cloud PubSub
- Google Cloud Functions

Data Preparation and Processing

- Google Cloud DataFlow

Processed Data Storage

- Google Cloud BigQuery

Data Visualization

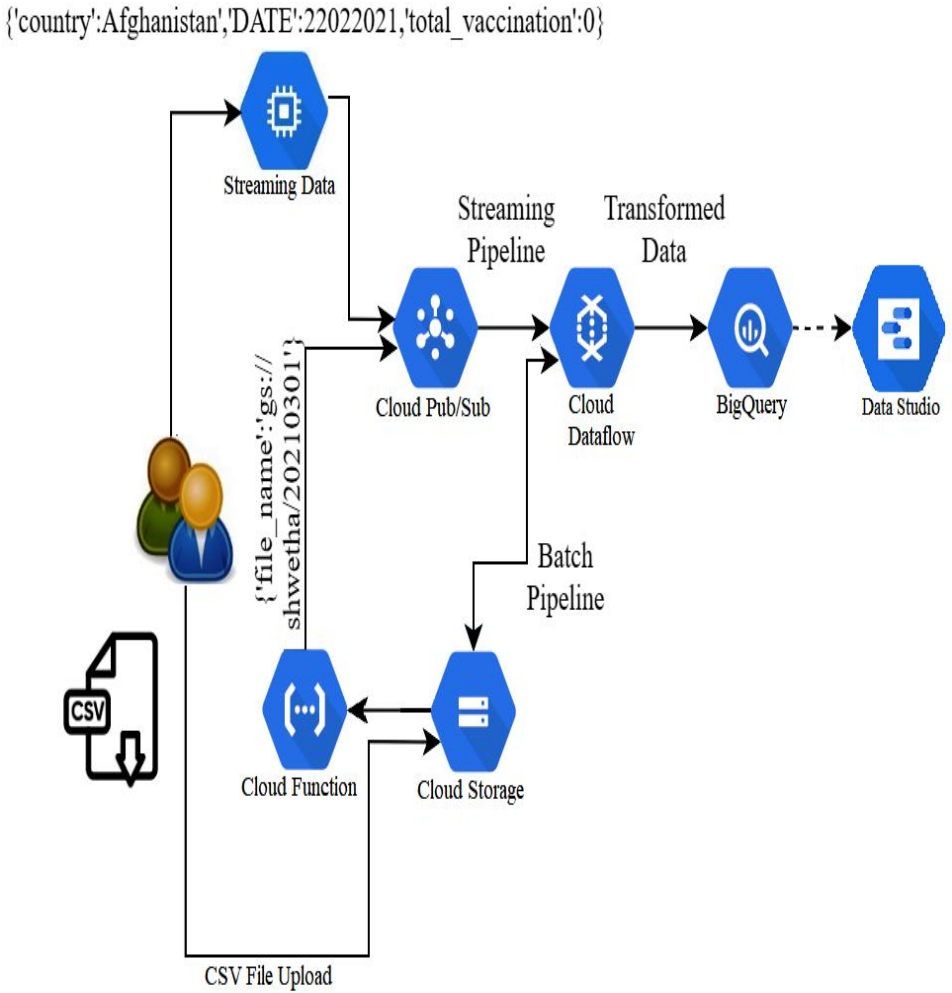
- Google Data Studio

Installation

- Install pip
- Set up python environment
- pip install apache-beam
- pip install google-cloud-storage
- pip install google-cloud-bigquery
- Pip install google-cloud-Pub/Sub

Visualizing Vaccination Progress Analytics Pipeline

- The streaming data is published to the **PubSub topic**.The **DataFlow** then reads the message from the Pubsub topic.
- The batch data is read by publishing the metadata to the **PubSub** topic using **Cloud Functions** which trigger the PubSub topic when the new csv file is added to the **Google Storage Bucket**.
- The **DataFlow** reads batch and streaming data from the **PubSub** topics
  and transforms the data into appropriate format.
- Once the merged data(Batch and Stream) is transformed, **Beam** will then connect to **BigQuery** and append the data to the designated table.
- The stored data in the **BigQuery** is then connect to **Data Studio** to carry out visualization.



Objectives

The main objective is to provide information about ongoing vaccine progress and find meaningful insights from the data.

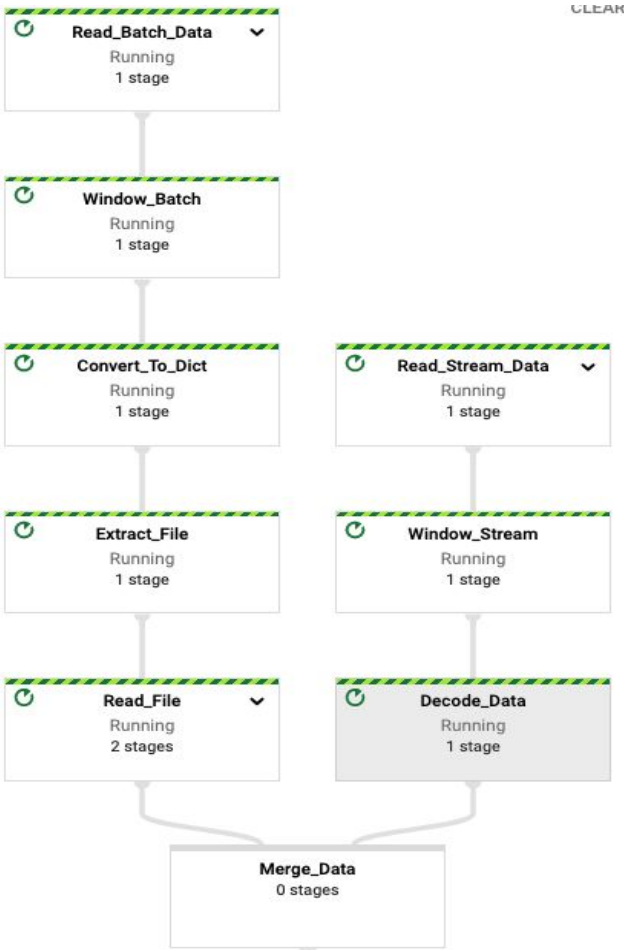
- Identify vaccines used by different countries and calculate the most popular vaccines in the world which gives us the vaccine that has the highest global reach.
- Compare vaccination trends of people who received at least one dose of COVID-19 vaccine with those people who are fully vaccinated.
- Find different categories of vaccines and understand which vaccine is being used in most of the countries around the world.
- To identify the country which has the highest number of vaccinated population.

Data Ingestion

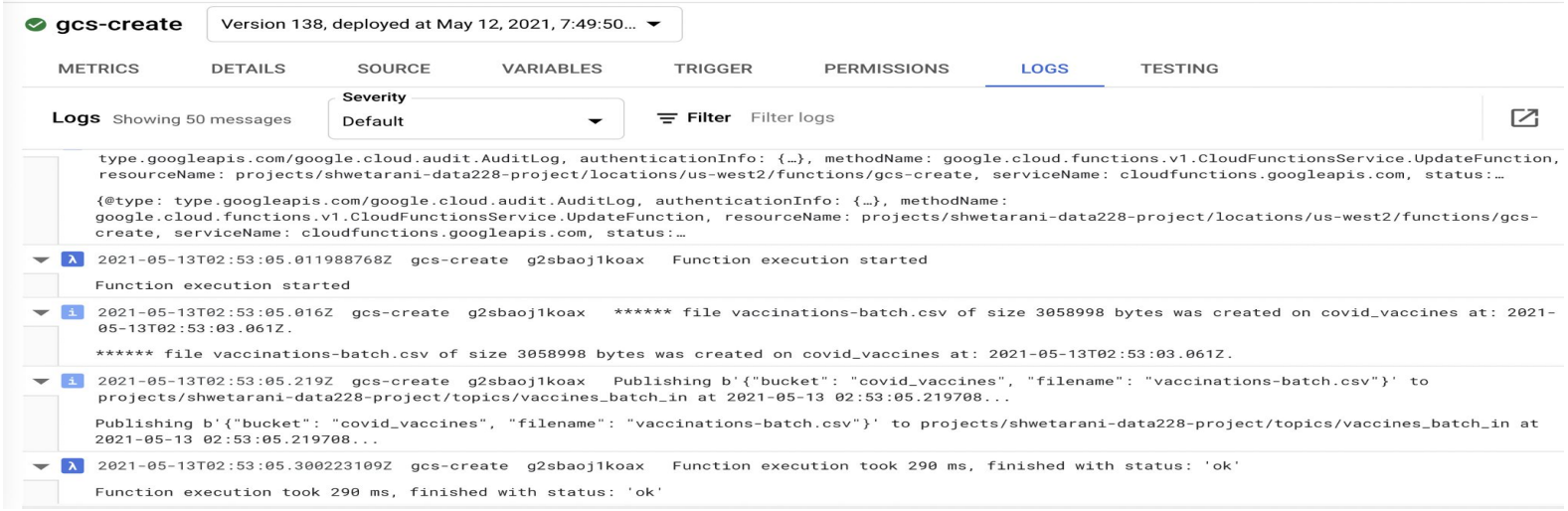
Data ingestion is the first stage to pull the batch as well as streaming data to perform processing.

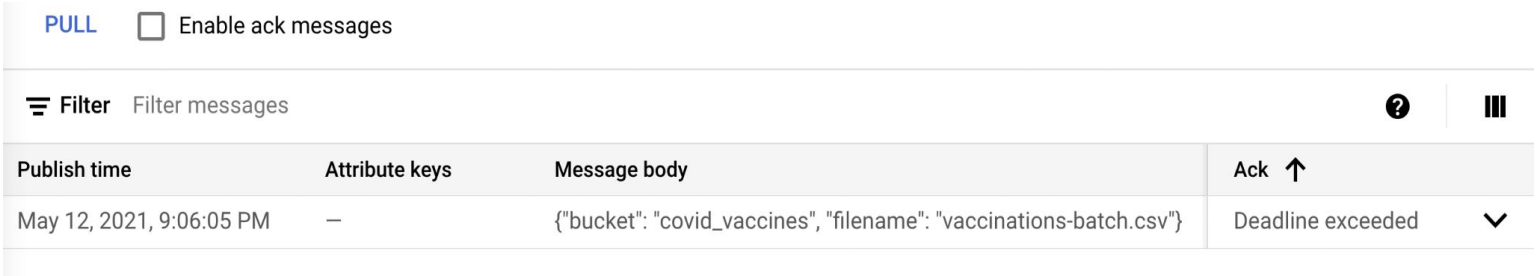
- Batch Data Ingestion is achieved by using Cloud Functions to trigger PubSub topic when the new csv file is added to the Google Storage Bucket.

- Streaming Data Ingestion is achieved by publishing continuously generated data to the PubSub topic.

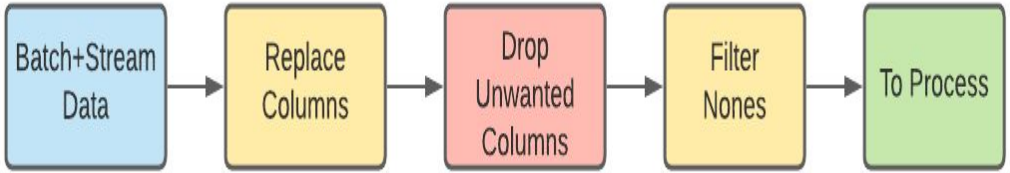


Cloud Function's Logs and Published metadata in the Pubsub topic



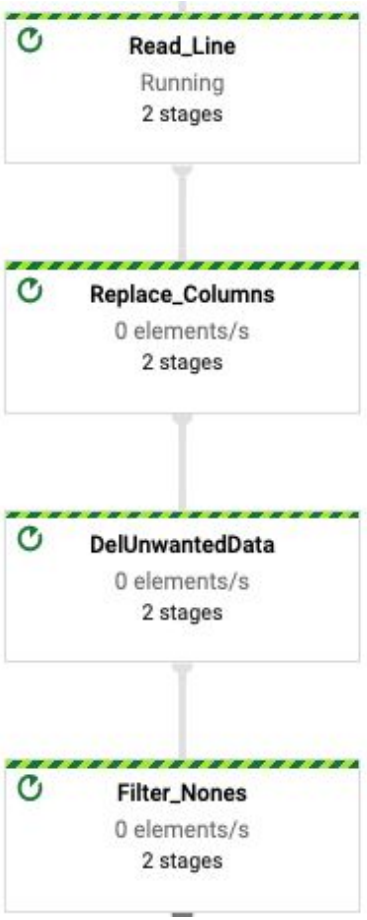


Data Preparation



Data preparation is the second stage, which is the task of blending, shaping and cleansing data to get it ready for analytics.

- Understanding the dataset structure to suit the purpose of analysis.
- Replacing missing values with **zero** for Float data type and **unknown** for object data type.
- Deleting unwanted columns which ended up as cleaned data.



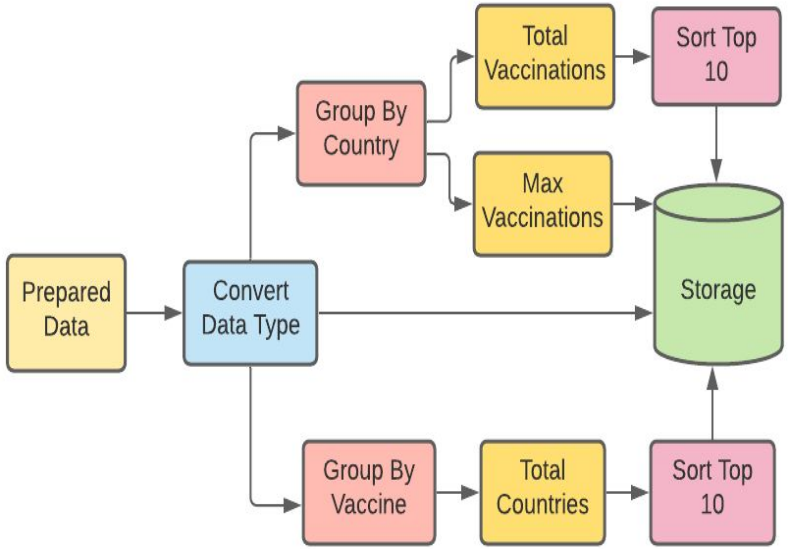
Data Processing/Analyzing

Data processing is the third stage, where the prepared data is collected and translated into usable information.

- As a part of our objectives we compare vaccination trend of total vaccinations per hundred,people vaccinated per hundred,people fully vaccinated per hundred by country.
- Identify vaccines used by different countries and calculate the most popular vaccines in the world.
- Identify most popular countries with highest vaccination.

Data Processing Cont'd..

- Prepared data is converted into appropriate data types for the further processing.
- Collecting fields by grouping by country and taking the sum of Daily vaccinations to calculate the most popular countries with highest vaccinations and sort top 10 countries.
- Finding maximum vaccination rates grouped by country.
- Collecting Fields by grouping by Vaccines and Counting the number of countries per Vaccine and sort top 10 Covid vaccines in the world.
- Then, writing the output to **BigQuery.**

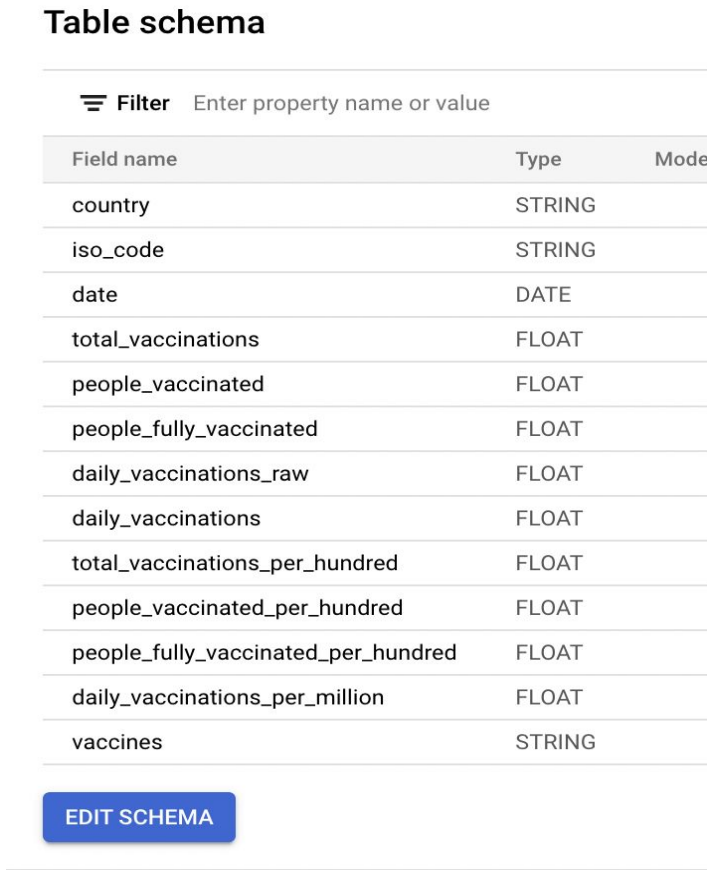


Transforms used in the pipeline

- **beam.io.ReadFromPubSub** : Reads the batch & streaming data from pubsub into the PCollection.
- **beam.Map:** The Map will accept a function that returns a one element for every input element in the PCollection.
- **beam.window.FixedWindows:**Applied window which divides up the data into fixed-width time intervals.
- **beam.GroupByKey:**Accepts keyed collection of elements and return collection with all values associated with that key.
- **beam.CoGroupByKey:**Accepts a dictionary of named keyed PCollections, and returns elements joined by their key.
- **beam.FlatMap:**Accepts function that returns a list where each of list elements is an element of the resulting PCollection.
- **beam.Flatten**:Merges two PCollection objects into a single PCollection object
- **beam.io.WriteToBigQuery** :Write transform to a BigQuery accepts PCollections of dictionaries.

Data storage after processing

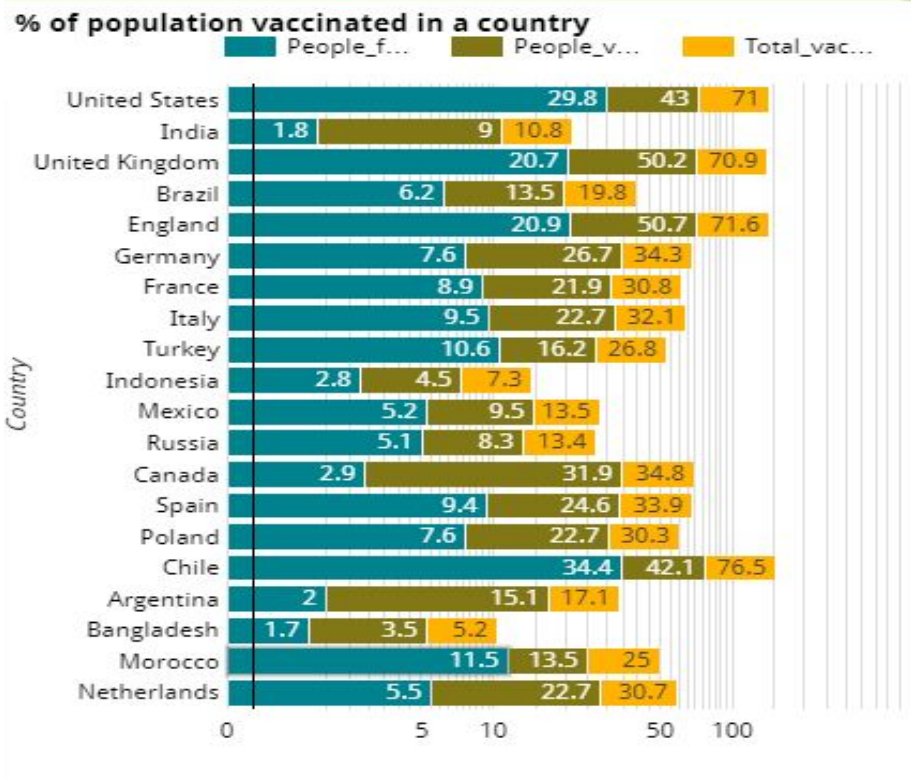
- Bigquery can handle large amounts of data storage
- It can process millions of rows in seconds.
- The large scale processed data is stored in BigQuery to carry out Exploration & Visualization using Data Studio.
- The table is created with the appropriate data types in order to store the data which is processed in Google Data flow.

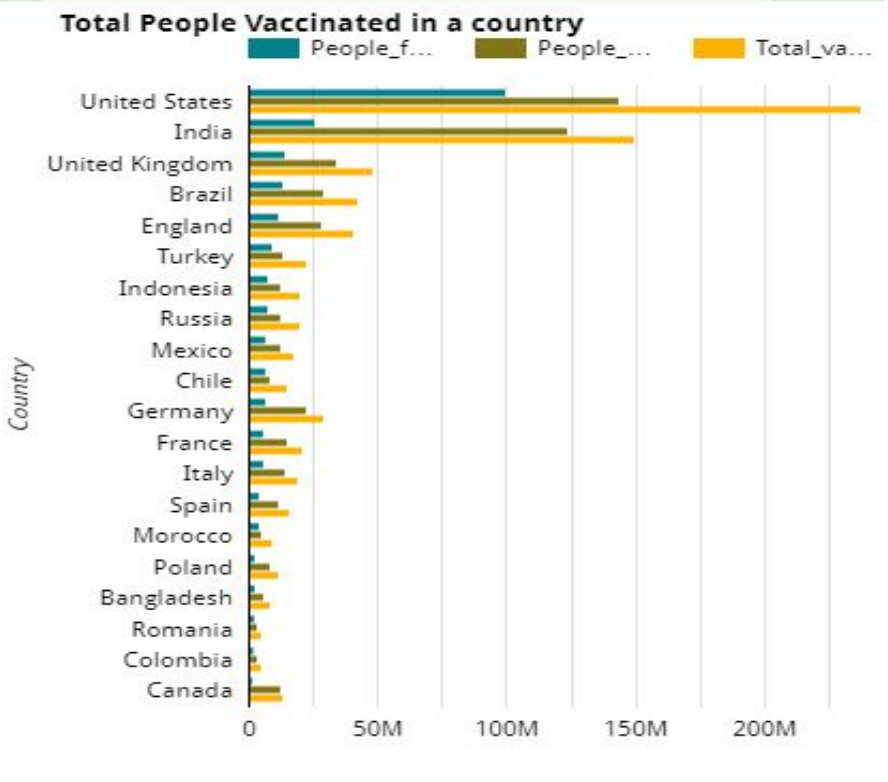


Data Visualization

- Data visualization is the practice of translating information into a visual context. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.
- In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.
- For our project we choose **GOOGLE DATA STUDIO** for visualizing our data.
- To visualize our **COVID-19 World Vaccination dataset**, the streaming and batch data that we processed using the pipeline will be stored in the BigQuery database. We have connected the database of BigQuery to the Data Studio as a data source and created a 15 minutes Data freshnes. With this, our dashboards will be updated automatically every 15 minutes with respective to streaming data.

COVID-19 World Vaccination Progress

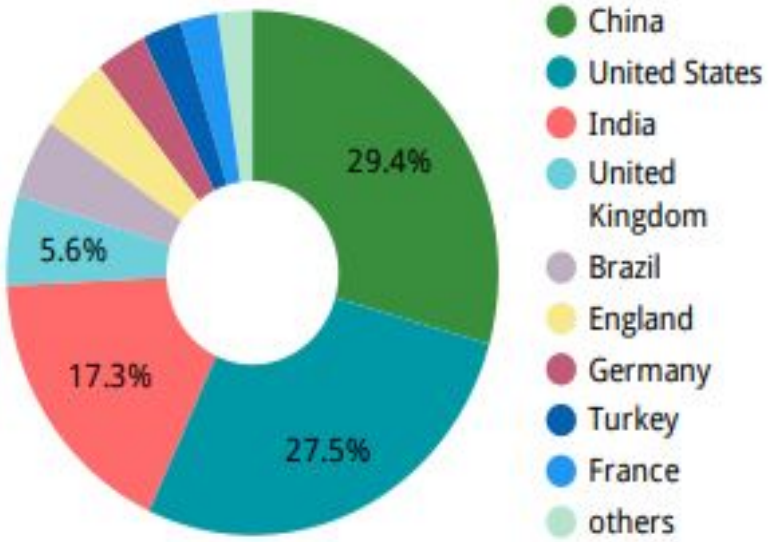




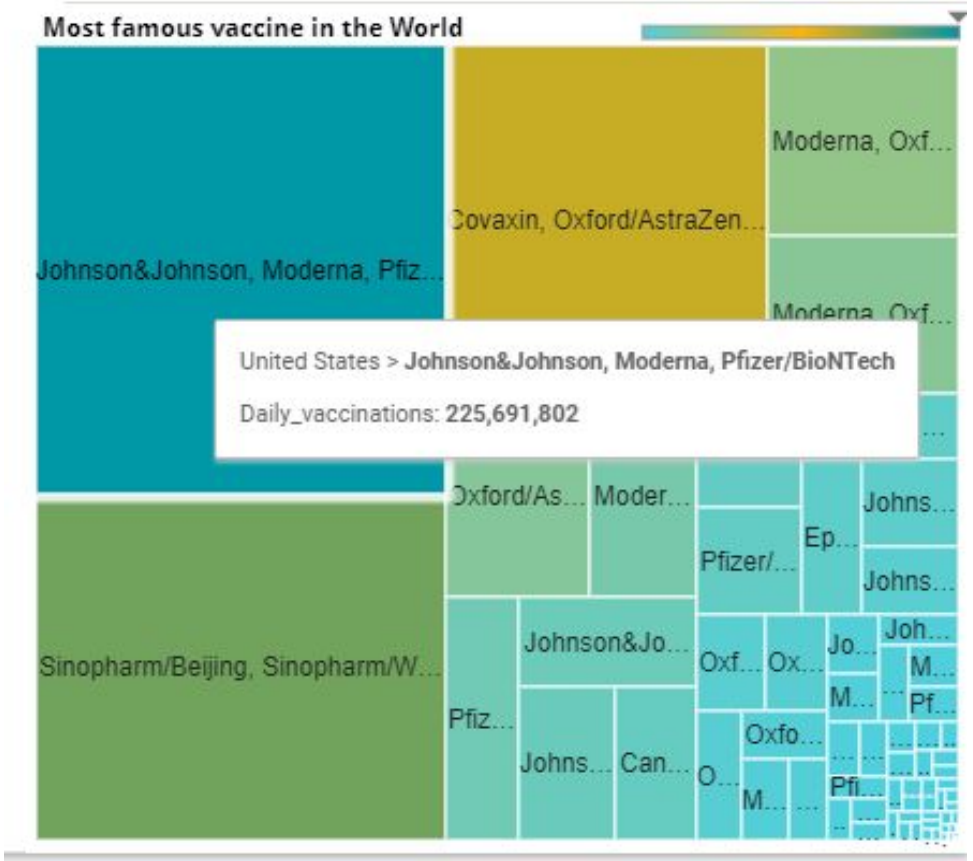
Dimensions: Country
Metrics: People fully vaccinated per hundred, people vaccinated per hundred and Total vaccinations per hundred.

Dimensions: Country
Metrics: People fully vaccinated, people vaccinated and Total vaccinations.

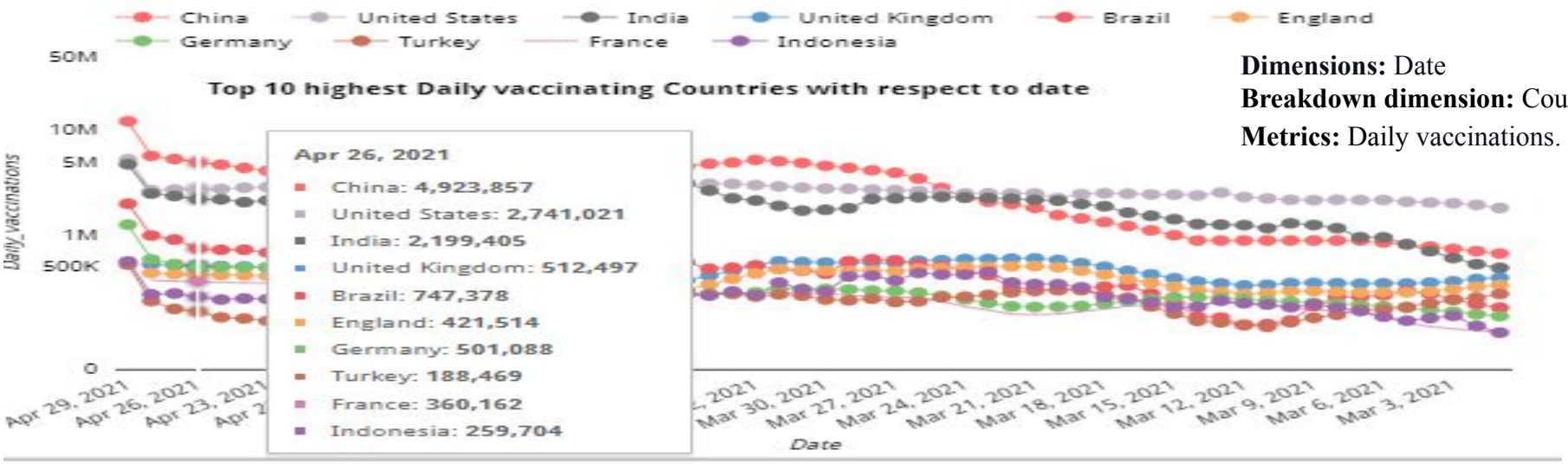
Country-wise contribution out of total vaccinations in the World



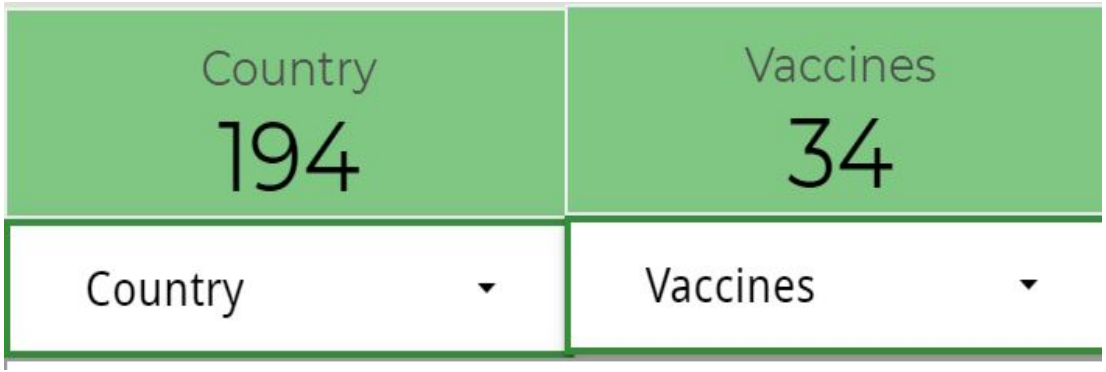
Dimensions: Country
Metrics: Total vaccinations.
Optional Metrics: People vaccinated, People fully vaccinated.



Dimensions: Country and vaccines
Metrics: Daily vaccinations.



Dimensions: Date
Breakdown dimension: County
Metrics: Daily vaccinations.



Scorecard: Total number of participating countries and Total number of available Vaccines

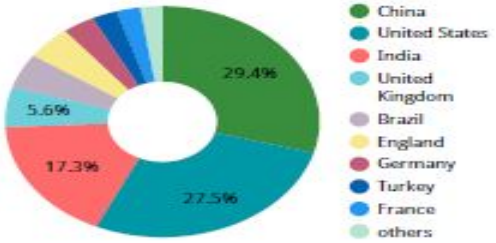
Control charts: Controlling fields are Country and Vaccines respectively.

COVID-19 World Vaccination Progress

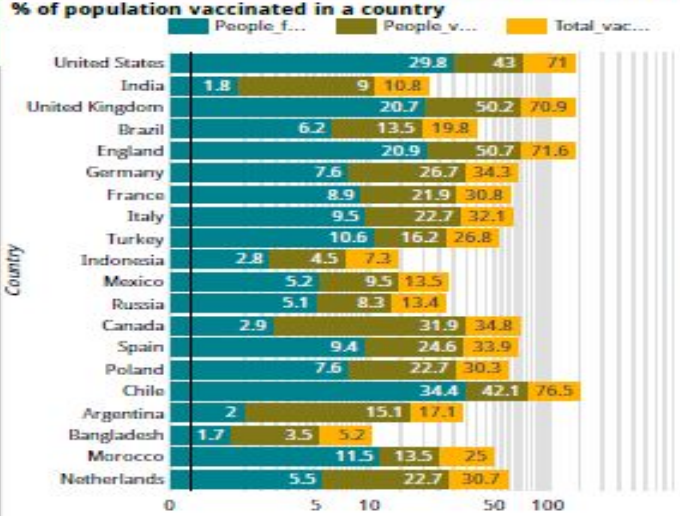
Country: 194 | Vaccines: 34

Country | Vaccines

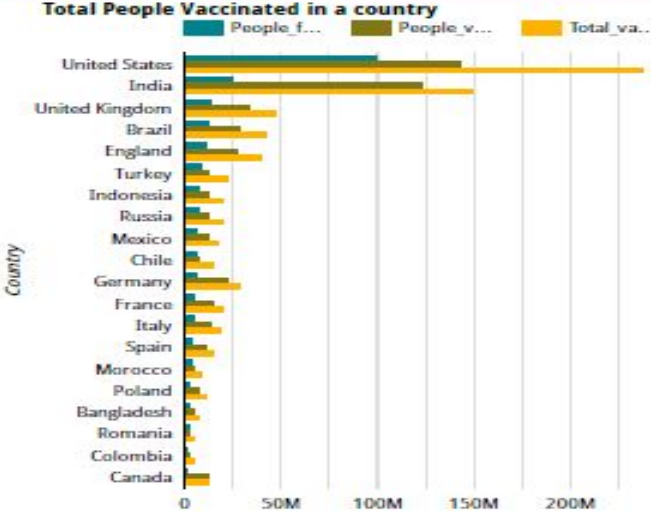
Country-wise contribution out of total vaccinations in the World



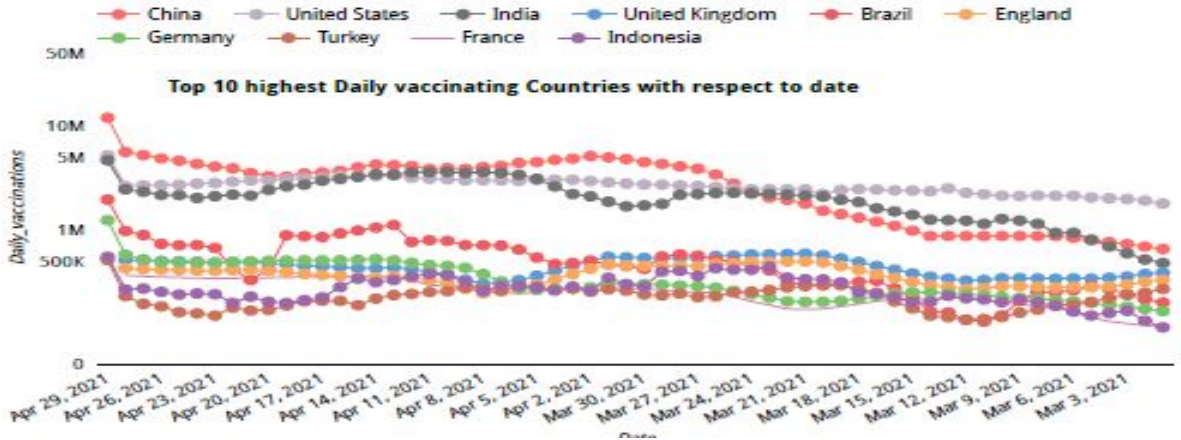
% of population vaccinated in a country



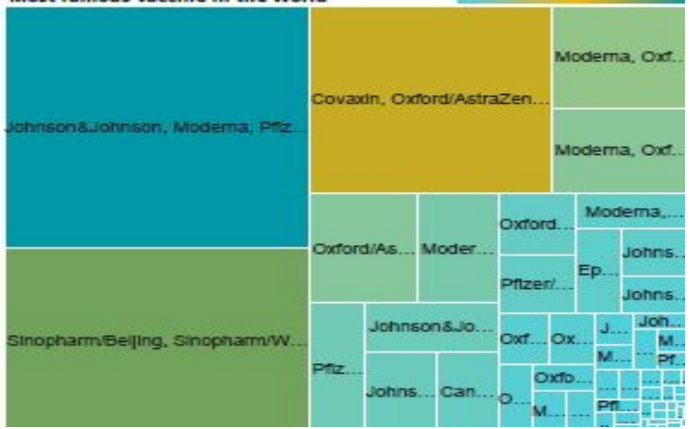
Total People Vaccinated in a country



Top 10 highest Daily vaccinating Countries with respect to date



Most famous vaccine in the World



Conclusion

- We have analyzed different ongoing vaccination programs by building an end-to-end data analytics pipeline using different google cloud platform services that involve ingestion, preparation, storage, and visualization.
- That is most useful in this pandemic situation to understand and analyze the vaccination stages and performance of the countries in the world. Every day, we are getting data from 194 countries as streaming data.
- We have stored world vaccination data from Dec 2020 to till today as batch data. So we will keep analyzing the data and provide our analysis in the form of visualization. That will be available through our Data Studio Dashboard.
- Our analysis for Covid- 19 vaccination can be used by the general public as well as governments for their performance analysis.

FutureWork

- If we have a data of total deaths caused by COVID 19 for each country that can be compared with the total vaccination to see if there is any inverse correlation. That is, if there are lower number of deaths in the country with the higher vaccinations.
- Forecasting vaccination trends using Machine learning prediction model.

Thank You