Github Link:

https://github.com/jimmy-nguyen-data-science/SF-COVID19-Vaccinations-pipeline

Data Pipeline for San Francisco COVID-19 Vaccinations

Design Document

Jimmy Nguyen, Yi Wang, Abby Tan

University of San Diego

Master of Science, Applied Data Science

February 21, 2022

Table of Contents

Project Objective

Background

On January 20, 2020, CDC confirmed the first COVD-19 case in the United States (CDC, 2022). The outbreak has continuously affected our communities and businesses. Since December 11, 2020, the two dose COVID-19 Vaccine from Pfizer-BioNTech has been approved and available for the public (U.S. FDA, 2021). The first vaccination in San Francisco was recorded by the City and County of San Francisco on December 12, 2022. Afterward, the two-dose vaccine from Moderna and the single-shot from Johnson & Johnson have become available as well.

General Overview

In this project, we designed an Extract-Load-Transform (ELT) data pipeline to extract San Francisco's COVID-19 vaccination daily updated data from a public data source named DataSF, load and transform the daily data into a data warehouse with the final output displaying an informative dashboard. The data pipeline needs data input as a source, and fortunately, DataSF provided developers with an API to work with in order to pull data. Therefore, we deploy a Python Script using Google's Cloud Scheduler and Cloud Functions to extract the data by calling DataSF's API. Then the data will be loaded into Google's data warehouse, BigQuery. Data transformations were executed by running SQL queries in dbt Cloud. Lastly, the clean and transformed data from dbt will be uploaded into BigQuery as new tables. These new tables populate the visuals for the final dashboard using Google's Data Studio. This design document describes the high-level overview for each component of the data pipeline and allows users to keep track of the necessary architecture information. Other Google's services will be discussed further in-depth in the following sections below.

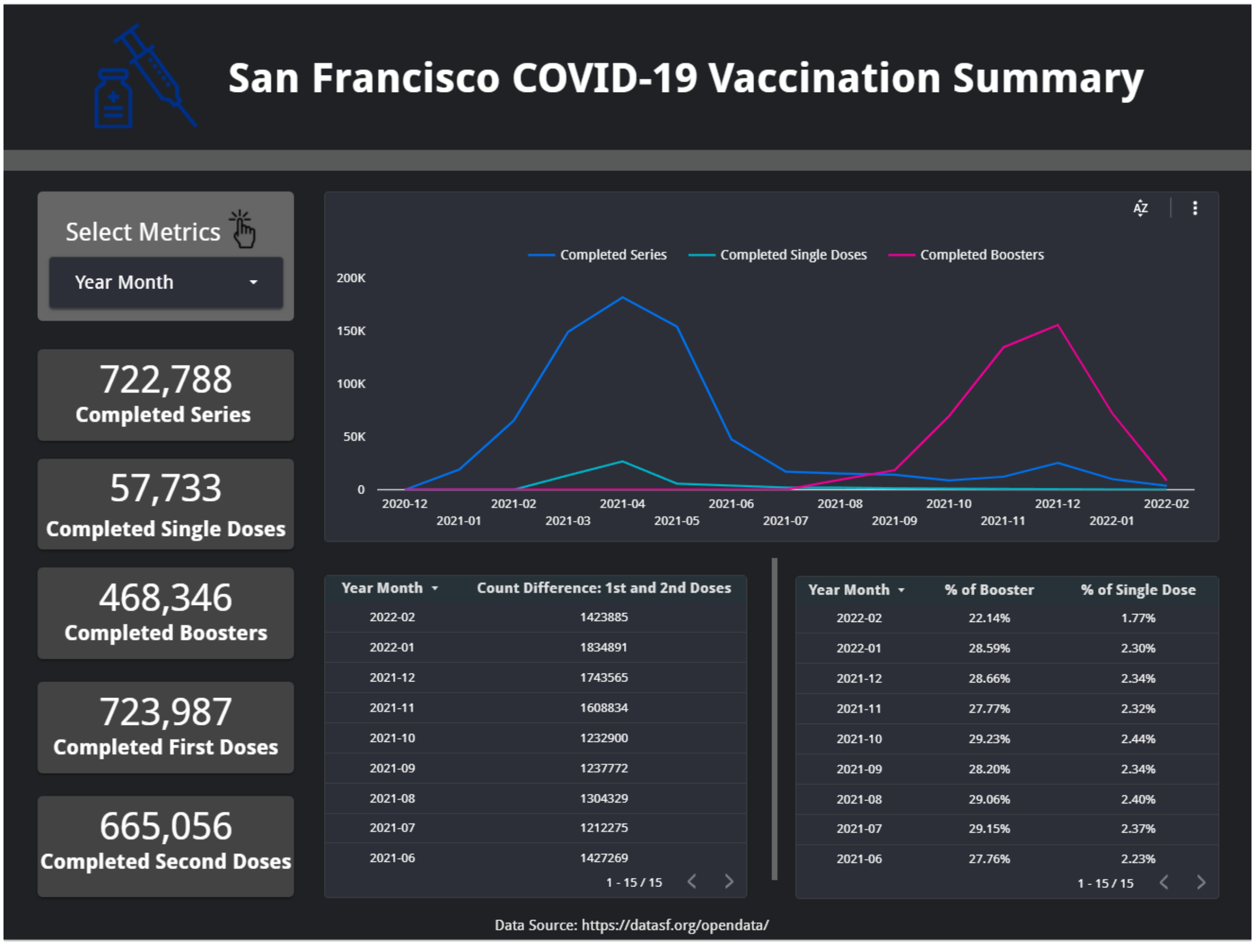
Data Source

Data Description

DataSF is an open source data sharing site to allow various personnel to utilize the city data of San Francisco. According to the data description, the COVID-19 Vaccination data is sourced from the California Immunization Registry (CAIR), run by the California Department of Public Health (CDPH) (DataSF). The data set was pulled in as a JSON file, and after converting it to a Pandas data frame, it has 15 columns and 441 rows so far, with the earliest record from December 21, 2022, and the most up to date record of February 25, 2022. The data size will be increased by one row each day after the data is updated. Each row is one day that represents the total count of each column. For example, the "date_administered" column has a value of February 24, 2022, will have total counts of the number of recipients who received their first dose, or second dose, or their only single dose, or if they have completed their series of doses. This information will be useful for us as we come up with different ways to utilize the data, such as aggregations by monthly or comparisons between 2020 vs. 2021. Such transformations will be done later by using dbt and visualizing in Google's Data Studio. Thus, data during the data ingestion stage are the raw data stored in Google's BigQuery data warehouse.

Pipeline Output

As shown in Figure 1, the final output of the pipeline is an interactive dashboard built from Google Data Studio. Data Studio is a free business intelligence tool to effectively integrate the data and allow users to generate reports, visualizations, Etc. While we use Google BigQuery as our data warehouse, the data can be easily streamed to Data Studio via Google Connector by simply identifying the project name, dataset, and table from BigQuery. With our pipeline output, the dashboard is pulling the transformed tables from the BigQuery named "dim_monthly_agg", "dim_monthly_cumulative", and "dim_monthly_boosters". The merged dataset called "comp_series_n_booster" utilizes the blended data feature from Data Studio to combine the tables of "dim_monthly_agg" and "dim_monthly_boosters". The dashboard visualizes the high-level summary of the San Francisco COVID-19 vaccination stats. By highlighting the key numbers from the dashboard, audiences can quickly and intuitively access the daily-updated San Francisco vaccination information.



San Francisco COVID-19 Vaccination Summary

Select Metrics

Year Month

722,788
Completed Series

57,733
Completed Single Doses

468,346
Completed Boosters

723,987
Completed First Doses

665,056
Completed Second Doses

Year Month	Count Difference: 1st and 2nd Doses
2022-02	1423885
2022-01	1834891
2021-12	1743565
2021-11	1608834
2021-10	1232900
2021-09	1237772
2021-08	1304329
2021-07	1212275
2021-06	1427269

1 - 15 / 15

Year Month	% of Booster	% of Single Dose
2022-02	22.14%	1.77%
2022-01	28.59%	2.30%
2021-12	28.66%	2.34%
2021-11	27.77%	2.32%
2021-10	29.23%	2.44%
2021-09	28.20%	2.34%
2021-08	29.06%	2.40%
2021-07	29.15%	2.37%
2021-06	27.76%	2.23%

1 - 15 / 15

Data Source: https://datasf.org/opendata/

Figure 1.

Dashboard of San Francisco COVID-19 Vaccination Summary

Note. This figure shows the output of the data pipeline from Data Studio.

Architecture Diagram

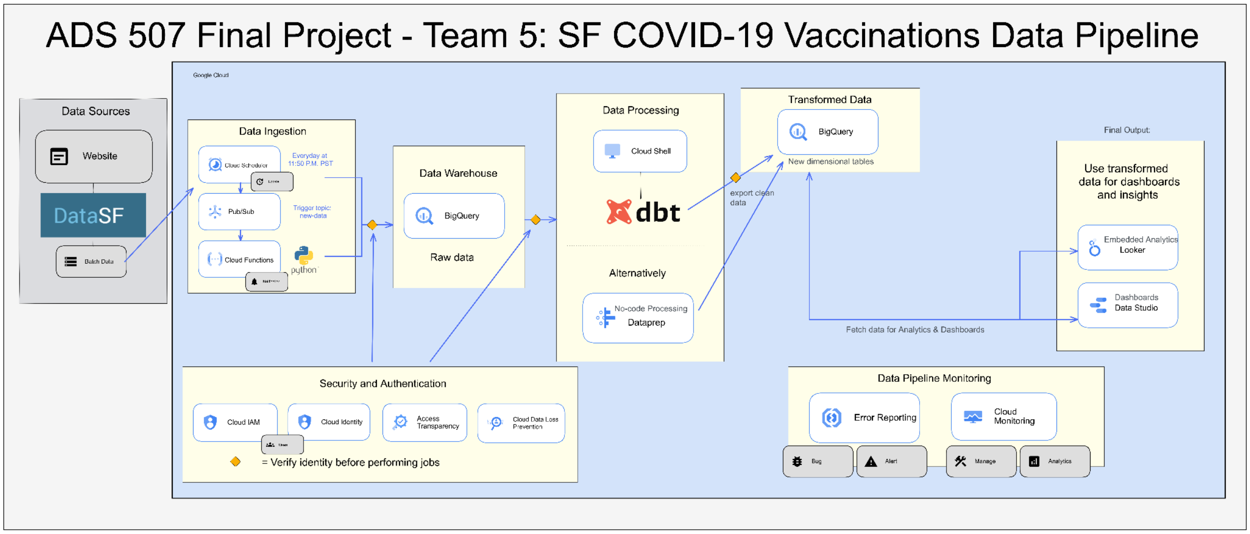


Figure 2.

San Francisco COVID-19 Vaccinations Data Pipeline

Note. This figure shows the data pipeline of the data is extracted, loaded, and transformed using Google Cloud data stack.

Service

Data Ingestion

Figure 2 shows the pipeline for our final project. The data pipeline will utilize Google's Cloud Scheduler and Cloud Functions to pull in the daily data as batch data and update an existing table in BigQuery. This works by having two distinct Python scripts. The first Python script is called "upload_starting_data" and is deployed once on Cloud Functions to create a schema and table called "daily-data" on BigQuery and load all the current data from DataSF's API. The second Python script is called "update_daily_data" and the purpose of this script is to append a new row in the existing table in BigQuery that represents a day's worth of vaccinations data. This deployment is scheduled by Google's Cloud Scheduler to run every day at 11:50 P.M. PST. It will send out a Pub/Sub message to trigger a topic called "new-data" and the Python script "update_daily_data" is programmed to only deploy from this topic. After this Python script is deployed, the new daily data will be appended as a new row into the "daily-table" table on BigQuery. The core components of data ingestion are Google Cloud Scheduler, Cloud Functions, and BigQuery.

Security and Authentication

However, in order to even ingest the data, there are security and authentication set in place before anyone can access it. Google's Cloud IAM and Cloud Identity need to be configured first in order to let certain users have BigQuery admin privileges. The scripts will require the user to read their authentication information as a JSON file when deploying on the command line in their local environment. Otherwise, when deploying the Python scripts on Cloud Functions, Google's IAM and Access Transparency can approve access to users through their Gmail accounts. Access Transparency also offers insight from near real-time logs when users access the data. This component of the pipeline provides great visibility and approval controls. Lastly, as part of the security and authentication of the pipeline, we also utilize a Google Cloud service called Cloud Data Loss Prevention (DLP). This service lets us gain visibility on the data we store and process, protect sensitive data with automated sensitive data discovery, and monitor our data on or off the cloud. Cloud DLP ultimately reduces the risks of a data leak and increases security.

Data Warehouse

The initial design of the pipeline included the priority to get the data and where to store the data. Ultimately, the final choice came down to Google's BigQuery data warehouse because the Google Cloud platform offers $300.00 of credit and 90 days free trial. In contrast, Amazon Web Services (AWS) only offers limited services for 30 days. As a start, BigQuery operates on a cloud platform, which means it has serverless architecture, allowing us to gather serverless insight and being able to scale our analytics automatically. Another attractive feature was logical data warehousing, meaning that we can process external data sources through BigQuery. This was the intention of using the highly modularized framework of data processing with dbt. Even though the data pipeline is only receiving one external data source currently, for future projects, Google's data transfer services allow us to work with other external sources such as Amazon S3. Data from multiple sources can easily be migrated or transferred into BigQuery automatically. The most beneficial and attractive benefits of BigQuery are the set-up and ease to use. BigQuery is easy and fast to set up in practically seconds. Once it is set up, we could query the data immediately to test the pipeline's operations further. This speed makes it easy to select BigQuery as the top choice for storing data. Lastly, BigQuery makes the process of loading and processing data simple since we only pay for what we use. This allows us to understand the data without the complications of building a data center. Google's BigQuery was the data warehouse that stored our daily pulled data from DataSF.

Data Processing

Now that raw data is stored in BigQuery, the next step in the pipeline is to process and clean this data. The data transformation tool of choice is dbt. This tool allows analysts or analytics engineers to write and perform data transformation tasks using SQL select statements. Simply put, the data pipeline follows an ELT framework where dbt is the ‘T’ or transformation in ELT. SQL queries in the dbt cloud can be written and compiled as models and materialized as views or dimensional tables in BigQuery. Such light transformations include simple aggregation by monthly or yearly data of San Francisco’s COVID-19 vaccination records of the total count. This is possible by breaking up one SQL select statement into multiple transformation models or SQL queries in order to have traceability in data lineage. The dbt tool also supports documentation and testing for quality checks and error reporting. Alternatively, since dbt only offers a 12 day free trial, Google Cloud also offers a service called Dataprep. Dataprep is a cloud data service tool that lets us clean and process data without code. Since it is a part of the Google Cloud platform, it is also serverless and easy to deploy and manage. Thus, Dataprep can be used as an alternative to replace dbt in the pipeline with its main feature for no-code processing.

Data Pipeline Monitoring

Data pipelines’ health and integrity need to be closely monitored. Otherwise, it will greatly impact performance, especially if a service fails. Google’s Error Reporting service helps us identify and understand when there is an error in the pipeline. For example, if an operation fails from one of the Python scripts in the Cloud Functions service, it will notify us immediately by text messages or emails. Not only that, but we can further understand the problems by observing the reports of error logs. This will come in the form of dashboards for aggregating different types of errors, including if something fails to execute from a recurring schedule. This will show a dedicated view of the error's time of failure, occurrences, and origin and display what incident we have resolved or are still unresolved. The Error Reporting service works in conjunction with Google's Cloud Monitoring to provide visibility in the performance and overall health of the infrastructure. These services create automated dashboards and alerts from system metrics and error logs to identify trends and prevent issues. Google's Error Reporting and Cloud Monitoring services are great for monitoring and integration with the current data pipeline.

Transformed Data and Final Output

After the raw data is fully processed and materialized as new tables in BigQuery, the final output of the pipeline is to use the transformed data to explore and extract insight in Google Data Studio. Google Data Studio allows analysts to create customizable informative dashboards or reports with internal or external sources. Since the transformed data comes from BigQuery, the dashboard will reflect the same source freshness from BigQuery as data is being updated and transformed daily. This service was chosen over dashboard visualization software such as Tableau because of its simplicity and automated connection. Figure 2 illustrates a dashboard as the final output of this data pipeline.

System Gaps

Meta has recently gone deeper into using large data volumes, for example, using billions of tagged Instagram images in ImageNet classification to achieve new record accuracy. This shows that increasing the training set size can improve model results even for problems with large, high-quality datasets. However, our system does not have an efficient way to control the dataset when it grows. We can train artificial intelligence models to manipulate the data set for the feature design when the dataset overgrows. Figure 2 shows that our system has exceptional data security and protection. For example, the system has Cloud IAM, Cloud Identity, Access Transparency, and Cloud Data Loss Prevention between data ingestion, data warehouse, and transformed data. Since our system is extensible, there are plenty of options for future migrations to different data warehouses such as Snowflake, Amazon Redshift, Azure Synapse, etc.

Reference

CDC Museum COVID-19 Timeline. (2022, January 5). Centers for Disease Control and Prevention. https://www.cdc.gov/museum/timeline/covid19.html#:%7E:text=January%2020%2C%20 2020%20CDC,18%20in%20Washington%20state

DataSF. (2021, February 11). *COVID vaccinations given to SF residents over time*. Retrieved February 25, 2022, from https://data.sfgov.org/COVID-19/COVID-Vaccinations-Given-to-SF-Residents-Over-Time/bqge-2y7k

U.S. Food and Drug Administration. (2021, August 23). *FDA Approves First COVID-19 Vaccine*. U.S. Food and Drug Administration. https://www.fda.gov/news-events/press-announcements/fda-approves-first-covid-19-vaccine