

Data Engineering 2022

Final Project

CovidWatch

LT9 | La Rosa • Soriano • Syllim



Table of Contents

Executive Summary	5
Introduction	6
Data Collection	7
Design Considerations	8
Data Architecture	9
Overview	9
OLTP Database	10
OLAP Database	12
NoSQL	13
Data Lake	14
API	15
Visualization	15
Conclusion	22
Recommendation	23
References	24
Appendix	25

I. Executive Summary

Despite the rising number of daily COVID-19 cases worldwide and the clear role that vaccines play in reducing mortality rates, vaccine hesitancy remains a persistent issue. Hence, having knowledge of people's sentiments towards vaccines in addition to relevant COVID infection statistics may just be what policymakers and health authorities need in determining necessary actions.

Given this need, we propose CovidWatch, an automated end-to-end database solution that extracts reliable data on COVID statistics and people's sentiments on a daily basis and provides policymakers and health analysts the means to derive meaningful and actionable insights through a visualization tool and an API endpoint.

CovidWatch currently has two main sources of data - Our World in Data (OWID) for COVID-related statistics and Twitter for obtaining social media data from which sentiments are extracted. Meanwhile, CovidWatch's architecture was designed with consideration to the system's scalability, evolvability and simplicity. Utilizing managed database solutions provided by AWS, CovidWatch has the following components:

- OLTP Database - although the RDS database is currently being updated daily, CovidWatch is envisioned to eventually draw from different sources in real-time, at which point updates will be done the moment new data (vaccination rates, cases, deaths, etc.) is received.
- OLAP Database - The Redshift database enables faster querying with its columnar format and allows dimensional analysis with its location and time dimension tables. This provides policymakers and stakeholders suitable statistics that apply to their jurisdiction and use case.
- NoSQL Database - the schemaless design of the NoSQL database enables preservation of all tweet information and provides maximum flexibility by allowing use cases beyond the sentiment analysis employed in this project.
- Data Lake - CovidWatch's S3 data lake, through its two zones, maximizes useability of the data it collects. The gold zone enables business analysts to easily extract meaningful insights from easy-to-use purpose-built data. Meanwhile, the landing zone preserves unaltered information for use by data scientists in advanced analytics and machine learning work.
- Scheduled ETL Jobs - DAGs implemented on Airflow perform the automated daily update of all the databases. For COVID data, DAGs download the data, perform ETL, and store the processed data on RDS, Redshift, and the S3 data lake . Meanwhile, another set of DAGs scrape tweets, perform sentiment analysis using Comprehend, and store results on both the S3 data lake and DynamoDB.
- Dashboard - A Quicksight dashboard which updates automatically with the data is provided to analysts to aid in decision making. This dashboard not only provides valuable COVID statistics and sentiment data in just a few clicks, but also allows for limited machine learning capability.
- API - Access is also provided to the general public through a REST API, enabling them to take advantage of the raw data for any value-adding use case.

Although CovidWatch serves its purpose of providing covid-related information, many improvements can still be made. Sentiment data can be expanded to include data from other sources like GDELT and Facebook posts. Meanwhile, COVID data can be more granular by sourcing information from individual sources (e.g. hospitals, cities, etc.). Despite these limitations, we firmly believe that CovidWatch remains to have the potential to improve the perception of vaccines and align policy with public opinion by providing accessibility of information surrounding covid and vaccines through the use of data engineering tools.

II. Introduction

The biggest effort to end the Covid-19 pandemic is the use of vaccines. To gain herd immunity, 70% of the population must be vaccinated. However, vaccine hesitancy is an issue policy makers need to address worldwide . In the Philippines alone, only 64.1% have been fully vaccinated as of June 12, 2022. The worldwide total is even less than that at only 61.3% (Our World In Data, 2022). Other studies have posed the question of marketing through the use of marketing theories to target certain consumer profiles, similar to marketing any new product entering the market. Social media can also be used in determining general sentiments toward vaccines and generating actionable insights for policy making especially in countries where it is widespread such as the Philippines (Cordero, 2022).

This study aims to answer the question: *“How might we eliminate vaccine hesitancy using multiple sources of data?”*

In this case, we expect a governing body similar to the World Health Organization, the Centers for Disease Control and Prevention or the Philippines’ Department of Health’s Inter-Agency Task Force who will be informing policies relating to the Covid-19 pandemic. This organization will rely mainly on Covid-19 data and statistics but will also use unconventional sources of data for public sentiments for their policy making.

III. Data Collection

A. COVID Data

Although the system is envisioned to query data from hospital and testing center databases in the future, the current implementation sources its data from the [Our World In Data \(OWID\) Github page](#). The variables collected include some of the most relevant COVID information like tests, cases, deaths, hospitalizations, and deaths. OWID sources their data from highly reliable sources including the Johns Hopkins Coronavirus Resource Center and health authorities of each nation (e.g. Department of Health for the Philippines). The tracked features and their descriptions as lifted from the source page are summarized in **Appendix A**.

The data is being retrieved through a single file download of the consolidated megafile which contains all of the variables being monitored. This is being done as new daily updates are not uploaded as single files, but are appended to the existing tables. As of June 2022, this daily update translates to a daily ingestion of 53.1MB for the system, but file size is expected to grow as more data is accumulated. However, this file size is small enough to be accommodated by the current process without any issues and will likely remain so for the near future unless major additions to the current table structure are made.

Albeit an admittedly convenient source of consolidated information, the OWID data remains sparse for some metrics, especially when it comes to developing nations (e.g. no vaccination information is available for the Philippines). Hence, the vision for CovidWatch is to eventually reach the point of collecting and aggregating piecemeal daily updates from different sources like hospitals and local government units.

B. Tweets

The current implementation of CovidWatch focuses on the sentiments of the people in the USA since they are one of the biggest countries in the world. Data was collected using a Python Library called [snsrape](#) that would scrape posts on social media sites, particularly Twitter. The team retrieved the most common hashtag used in the USA relating to the COVID-19 pandemic, such as #covid19usa. The data dictionary coming from Twitter API can be found on its [developer website](#).

The data was initially scraped from June 2020 to the present to match our COVID data. Then, a daily batch job would run to scrape the posts on Twitter the day before. Currently, the daily ingestion of tweets is around 400KB, with a performance of 50 posts per second. Still, file size is expected to grow enormously as the goal of CovidWatch is to understand the sentiments of all the countries and not only the USA.

IV. Design Considerations

In choosing the different technologies and components to employ in implementing the solution, we considered the following as the most important criteria in line with CovidWatch's mission:

Scalability

As stated in the previous section, the size of the ingested data is expected to rise as more countries and more data sources are accommodated. It is also highly conceivable for CovidWatch to eventually handle queries from hundreds or even thousands of different analysts who desire to make use of the valuable data it holds. Hence, the system must have viable means of increasing both its storage capacity and its throughput.

Evolvability

In addition to the ability to increase capacity where needed, the system also has to be flexible to cater to any previously unanticipated use cases. Hence, in addition to having the means to accommodate new processes, the data being collected should also be flexible enough to serve different purposes and the interfaces to access this diverse data should be available. For instance, the Twitter data being collected can be utilized for many other different use cases other than the sentiment analysis implemented currently.

Simplicity

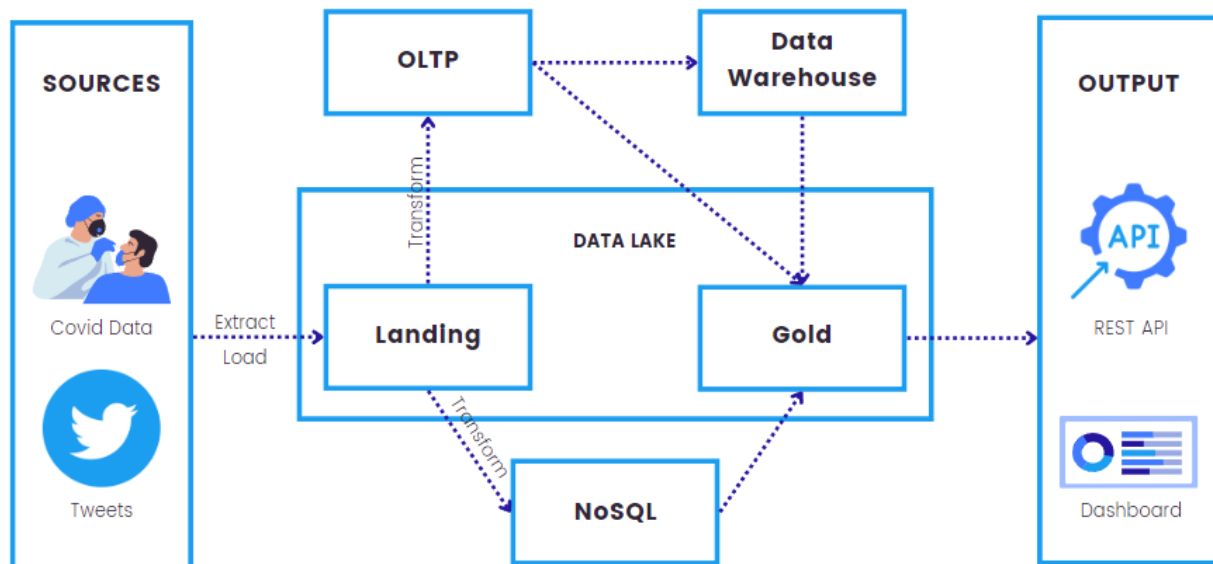
We hope for the system to be manageable by a small team of personnel who have minimal data engineering experience. Therefore, the addition of scale or adaptation to new features must not require highly specialized skills (e.g. data engineering, software engineering, network engineering, etc.).

V. Data Architecture

A. Overview

The figure below shows a high-level overview of the architecture we plan to implement to solve this use case. The team implemented an ELT (Extract-Load-Transform) pipeline. This is to ensure that raw copies of the files are kept for record keeping in the landing zone.

Figure 1. High-level overview of the architecture



The following components were chosen to satisfy each of the project's unique needs:

1. OLTP Database

The OLTP database for CovidWatch is envisioned to be updated daily with new data from various sources like hospitals, testing centers, and information from health organizations like each country's health authority (e.g. Department of Health or Ministry of Health) and WHO. Due to this anticipation for daily updates, the database was implemented through the Amazon Relational Database Service (or Amazon RDS) which is a managed relational database service. This allows for a complete single source of truth for clean Covid data.

2. OLAP Database

The OLAP database serves as the data source for the visualization application where policymakers, health authorities, and healthcare analysts can extract meaningful information.

3. NoSQL Database

Twitter data coming from the landing zone is expected to have an enormous size of content of at least 2 GB per day corresponding to the 240 countries that the CovidWatch plans to cover in the future. Hence, the need to use a NoSQL database to be able to scale out to multiple machines. A key-value pair NoSQL database was chosen to be able to analyze the Twitter daily data, having a more simple database. This can also be expanded if other forms of media will be digested.

4. Data Lake

The team implemented a data lake consisting of landing and gold zone for easier evolvability of the current and future use cases of CovidWatch. The NoSQL, OLTP and OLAP databases

will have a copy automatically updated in the gold zone. This would make the data available to future applications and improve models from newly cleaned data. It will also be used for data visualization to produce a report for stakeholders. The use of a data lake is also cheaper. We can start the databases only when new data is delivered for consumption.

5. API

An API was created to access the OLAP database and NoSQL data in the gold zone to make it available to the public to crowdsource for solutions and insights. This also allows for transparency and auditing for the governing body because models and insights would be reproducible.

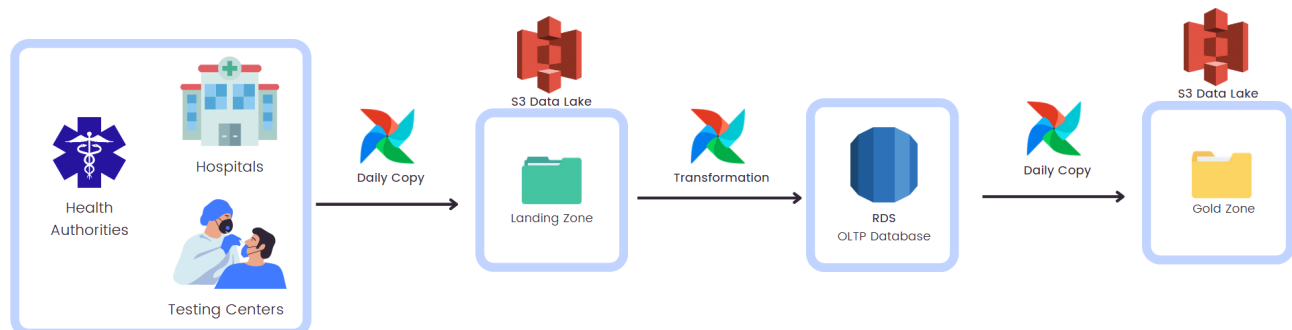
6. Visualization Tool

Amazon Quicksight was used for quick and easy implementation of a dashboard which provided basic statistics for flexible reporting to the stakeholders. This also allowed for basic machine learning implementations within AWS services such as forecasting and anomaly detection.

B. OLTP Database

The OLTP database for CovidWatch is envisioned to be updated daily with new data from various sources like hospitals, testing centers, and information from health organizations like each country's health authority (e.g. Department of Health or Ministry of Health) and WHO. Due to this anticipation for daily updates, the database was implemented through the Amazon Relational Database Service (or Amazon RDS) which is a managed relational database service.

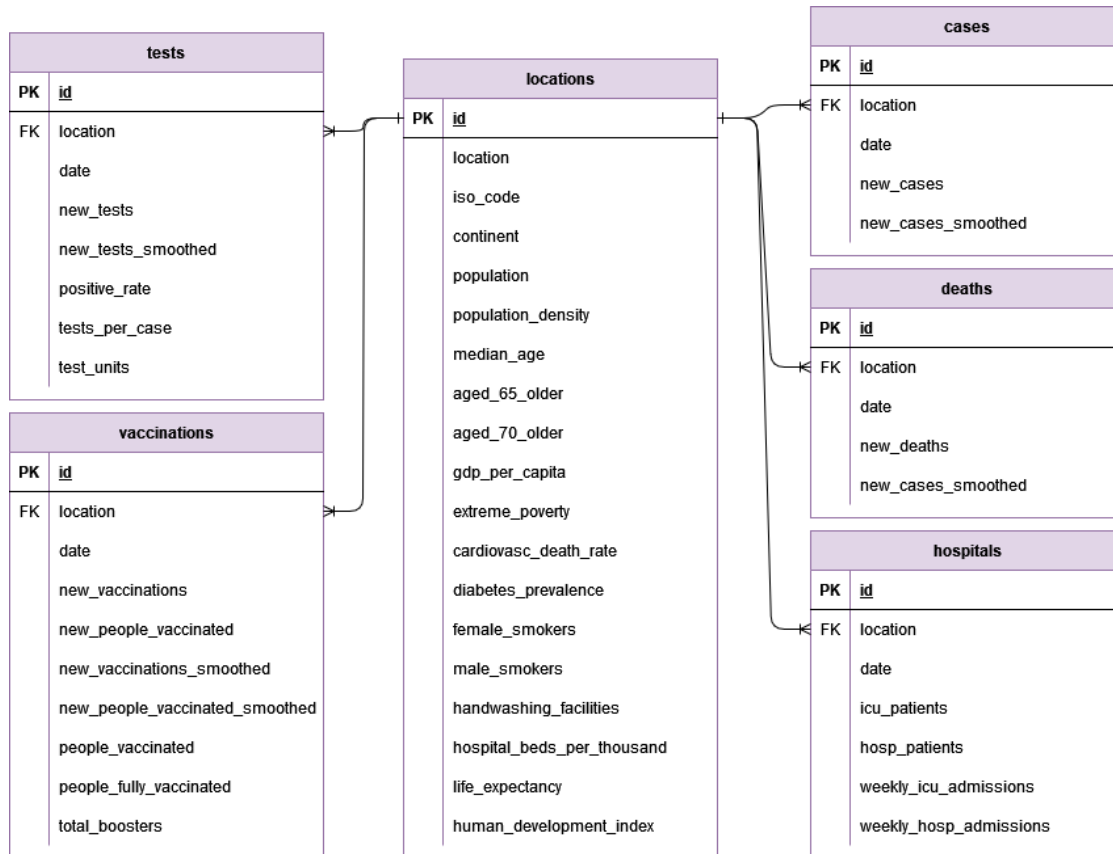
Figure 2. OLTP Pipeline



The current implementation copies data from OWID to the landing zone of an Amazon S3 data lake using a scheduled DAG. From the S3 folder, data is split into different columns, processed, and inserted into the OLTP database through a scheduled DAG implemented through Amazon Airflow. The DAG script for data extraction, storage to S3, and processing for RDS storage can be found as the `copy_step` and `insert_to_rds_step` steps in the attached file `extract_covid_data.py`.

Not all of the OWID data is updated on a daily basis. Some information is updated weekly (e.g. hospital admissions) while others are updated once a month (e.g. mortality), more than 1 month after the actual dates. Hence, to help minimize the required writes to the database while ensuring validity of the data, only the last 2 months of the RDS data are updated by the DAG. This is done by deleting all information for the past two months and inserting the new data. The OLTP database schema is shown on Figure 3.

Figure 3. OLTP Schema



The schema contains the following tables, each expected to draw from different sources of data:

1. locations

This table contains location information relevant to managing COVID cases including population density, proportion of elderly, cardiovascular death rate, and number of hospital beds. While the current implementation only contains information on a country level, the hope is to eventually have granularity at the level of cities or regions.

2. tests

This table contains daily test numbers and positivity rates, information that is relevant in determining any additional COVID testing resources.

3. vaccinations

Daily vaccinations is currently one of the best pieces of information that help prevent the spread of COVID, hence information like people vaccinated and total boosters administered are also provided where available.

4. cases

The cases table contains the number of new cases per day which helps determine needed alarm levels and precautions. Smoothed data is also provided where some days are skipped.

5. deaths

Like cases, the deaths table contains daily numbers which are smoothed where some days are skipped. This is highly relevant in identifying hospital resources like ICU beds.

6. hospitals

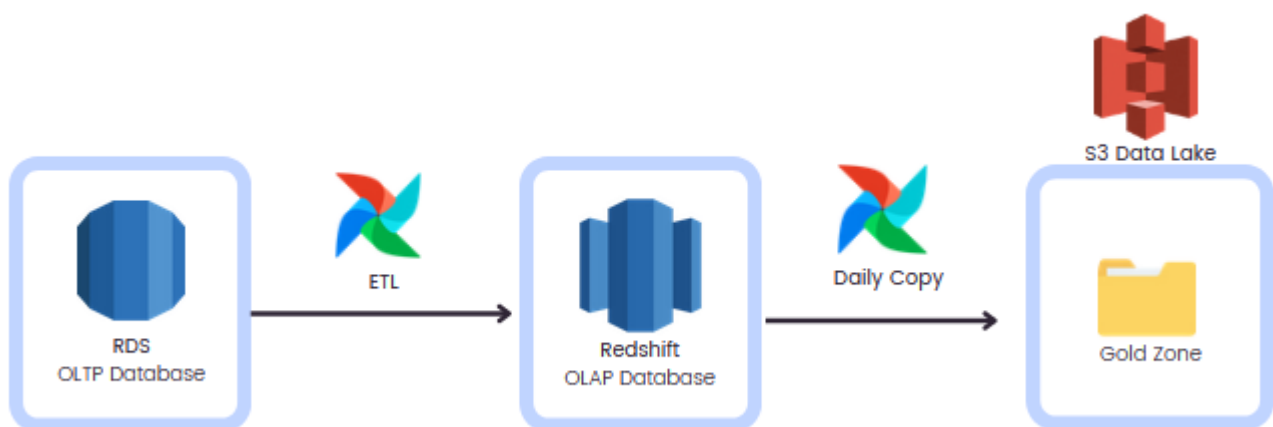
The hospitals table contains numbers for COVID-related admissions including ICU patients. Coupled with data on the country's available beds, this information is helpful in determining the need for emergency medical facilities.

C. OLAP Database

The OLAP database serves as the data source for the visualization application where policymakers, health authorities, and healthcare analysts can extract meaningful information. The current implementation draws data from the OLTP Database, transforms the data into facts and dimensions tables, and inserts them into an OLAP Database that was created using Amazon Redshift which was chosen because of its columnar nature. Queries to be made from the OLAP database are likely to involve only one column at a time over a defined timeline, hence this format enables faster query results.

This whole process is done through the use of DAGs via Amazon Airflow. The script for this process can be seen in the `insert_to_redshift` step of the same `extract_covid_data` DAG used for creating the OLTP database. As with the RDS updates, only the last two months of OLAP data is updated daily to reduce the necessary load on the system.

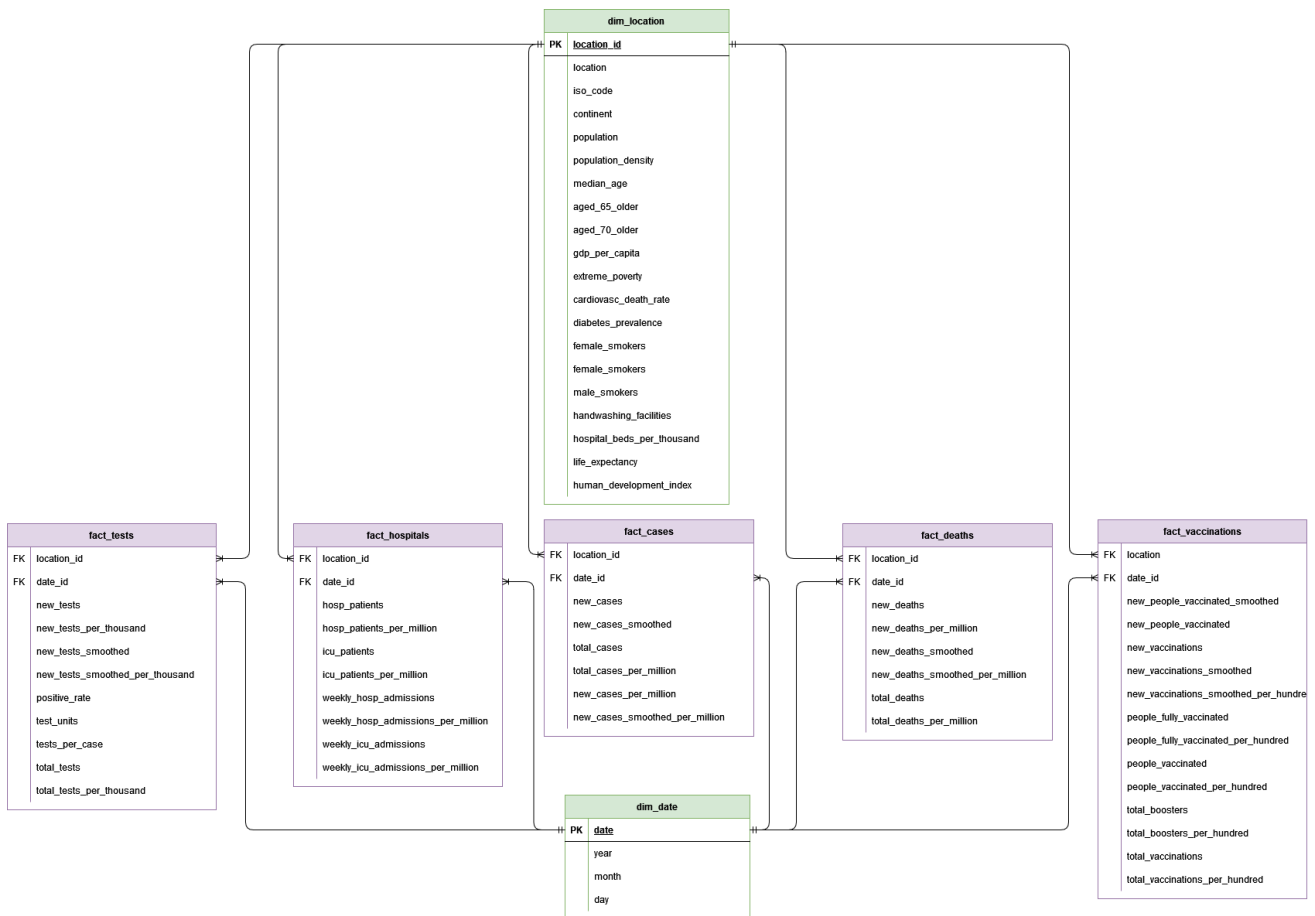
Figure 4. OLTP to OLAP Pipeline



The OLAP database contains 5 facts tables (`fact_cases`, `fact_tests`, `fact_hospitals`, `fact_deaths`, and `fact_vaccinations`) which, in addition to having aggregated information from corresponding tables in the OLTP, also contains numbers normalized based on the country's population. This information allows inter-country benchmarking by providing more comparable metrics. The fact tables remain separated by the relevant statistics they describe in order to avoid having the need to update whole rows just for one new piece of information affecting just one family of columns.

To provide grouping, labeling and filtering functions, two suitable dimensions were created - the locations dimension and the date dimension. These two dimensions allow analysts to focus on specific locations within specified timeframes to extract the most relevant information that would inform necessary policy and information drive measures within each jurisdiction. The schema, along with the features for each table, are summarized in Figure 5.

Figure 5. OLAP Schema



D. NoSQL

Figure 6. NoSQL database pipeline

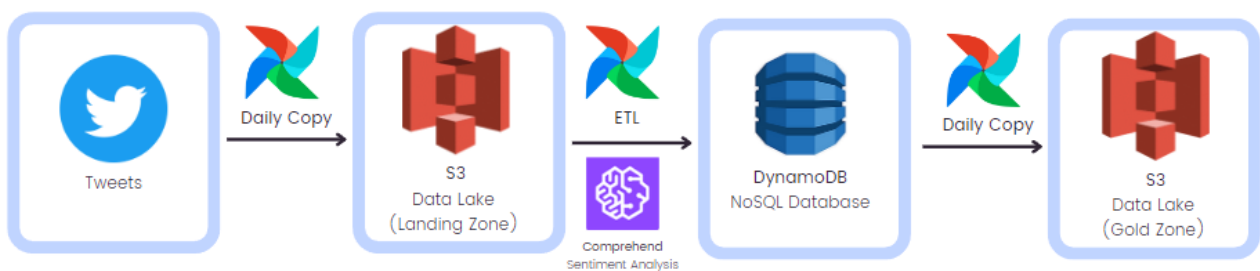


Figure 6 shows the pipeline for the NoSQL database to get the sentiments of the people. Airflow runs a daily job that would scrape posts from Twitter with hashtags relating to the COVID-19 pandemic. Then, the scraped data will be stored directly in the landing zone. Although content is the only important feature for sentiment analysis, the team opted to save all of the tweet features for future use such as network analysis. Some preprocessing was done to the content of the tweet, such as removing hyperlinks and hashtags. Then, Amazon Comprehend is used to get the sentiment scores of the tweet content. The following features were extracted after performing preprocessing, sentiment analysis, and selection of features:

Table 1. Transformed data

Feature	Description
id	tweet_id
date	Date extracted
content	Content of tweet
replyCount	Count of replies
retweetCount	Count of retweet
likeCount	Count of likes
quoteCount	Count of quotes
hashtags	Hashtags present in the content
Sentiment	Sentiment text [Positive, Negative, Neutral, Mixed]
SentimentScore.Positive	Probability of the sentiment being positive
SentimentScore.Negative	Probability of the sentiment being negative
SentimentScore.Neutral	Probability of the sentiment being neutral
SentimentScore.Mixed	Probability of the sentiment being mixed
tokenized_content	Tokenized content

The transformed data is stored in a DynamoDB. The date was chosen as the partition key to analyze daily tweets, and the sort key was chosen to be the tweet id to ensure the uniqueness of the primary key for each record. Currently, it has an average request latency of 33 ms. Finally, a copy of the DynamoDB in a JSON format is saved to the Gold zone for further analysis by the Data Scientist and Business Analysts.

E. Data Lake

The team created an Amazon S3 bucket with two directories, with each directory corresponding to a data lake zone, namely the landing and gold zone. The team has created two zones to cover the present and future use of the data. We have also identified three user groups: data engineer, data scientist, and business analyst that would work on some of the data lake zones. The description of each zone and its access control are discussed below:

1. Landing Zone

The landing zone contains the raw data coming directly from the data source which are Our World and Twitter data. This data lake zone was created for the possible use of data in the future which are currently not implemented in the CovidWatch dashboard. A sample use case could be network analysis of spread of disinformation regarding Covid-19 vaccines, and many more. The team has provided **full access to data engineers**, so that they can easily transform the data to the Gold Zone if needed.

2. Gold Zone

The gold zone contains the cleaned and processed data for data analytics and visualization. The output of the OLTP, OLAP, and NoSQL databases are located in this data lake zone which is the data used for CovidWatch in the present. Data is provided with full access to the **Data Scientist** and **Business Analysts**. Data scientists can use this data to forecast the possible number of vaccines and hospitalization or retrain the NLP model for sentiment analysis, among other use cases. Meanwhile, Business Analysts can look closely at the data provided for reporting to stakeholders.

F. API

Amazon provides an API service called API Gateway which can automatically connect to different API services. We specifically use this to connect to the data lake to access a copy of the OLAP database and a copy of the NoSQL database. Two resources were created, one for NoSQL and another for OLAP, each with a single GET method. The NoSQL database has a definite link, while the OLAP database files can be accessed by changing the file name at the end of the URL.

OLAP:

https://ep3yyvtgeb.execute-api.us-east-1.amazonaws.com/development/covidwatch/olap/dim_dates_000

```
date,year,month,day
2021-04-09,2021,4,9
2020-06-11,2020,6,11
2021-10-15,2021,10,15
2020-04-21,2020,4,21
2020-09-16,2020,9,16
```

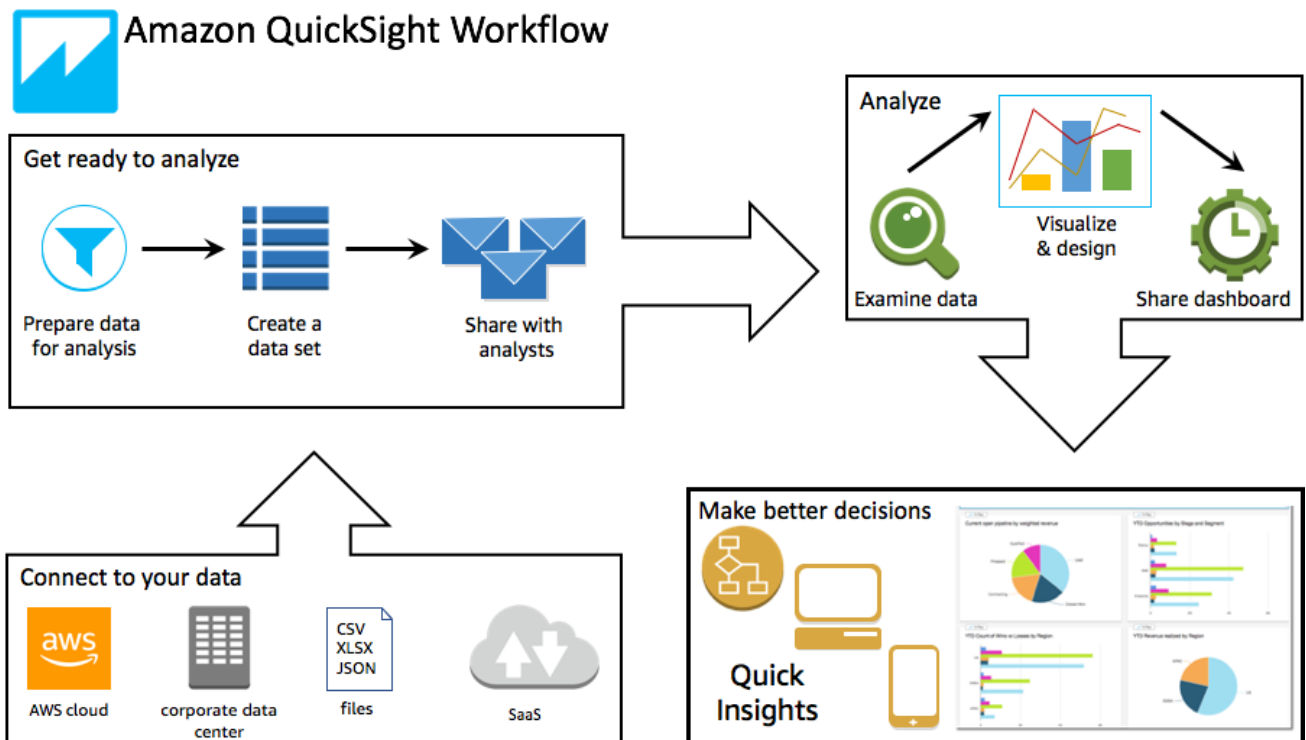
NoSQL: <https://ep3yyvtgeb.execute-api.us-east-1.amazonaws.com/development/covidwatch/tweets>

```
{"content": {"S": "RememberTheDead300k dead from Covid19USAWearAMask tcoEDK4EqUnZi"}, "retweetCount": {"N": "0"}, "hashtags": {"L": [{"S": "RememberTheDead"}, {"S": "Covid19USA"}, {"S": "WearAMask"}]}, "tokenized_content": {"L": [{"S": "RememberTheDead300k"}, {"S": "dead"}, {"S": "from"}, {"S": "Covid19USAWearAMask"}, {"S": "tcoEDK4EqUnZi"}]}, "SentimentScore.Negative": {"N": "0.03435024246573448"}, "Sentiment": {"S": "NEUTRAL"}, "SentimentScore.Positive": {"N": "0.005107102915644646"}, "likeCount": {"N": "0"}, "SentimentScore.Mixed": {"N": "0.0000466922835218534"}, "date": {"S": "2020-12-15"}, "SentimentScore.Neutral": {"N": "0.9604958891868591"}, "replyCount": {"N": "1"}, "id": {"N": "1338650481032376322"}, "quoteCount": {"N": "0"}}
```

G. Visualization

A dashboard was developed on Amazon Quicksight to aid in the decision-making of the relevant stakeholders. Amazon QuickSight is a secure, scalable and redundant business intelligence service provided by Amazon. The dashboard automatically updates as soon as the data sources change since it is connected directly to the data lake. It will also automatically inform the dashboard owner of any changes to the data source. The dashboard is broken down to seven sections: Overview, Philippine Cases, Vaccination in the US, Hospital Use, Sentiments, API, Feedback. Since this is just a proof of concept, we expect that these features would change with feedback from the stakeholders. Figure 7 shows the Amazon QuickSight workflow.

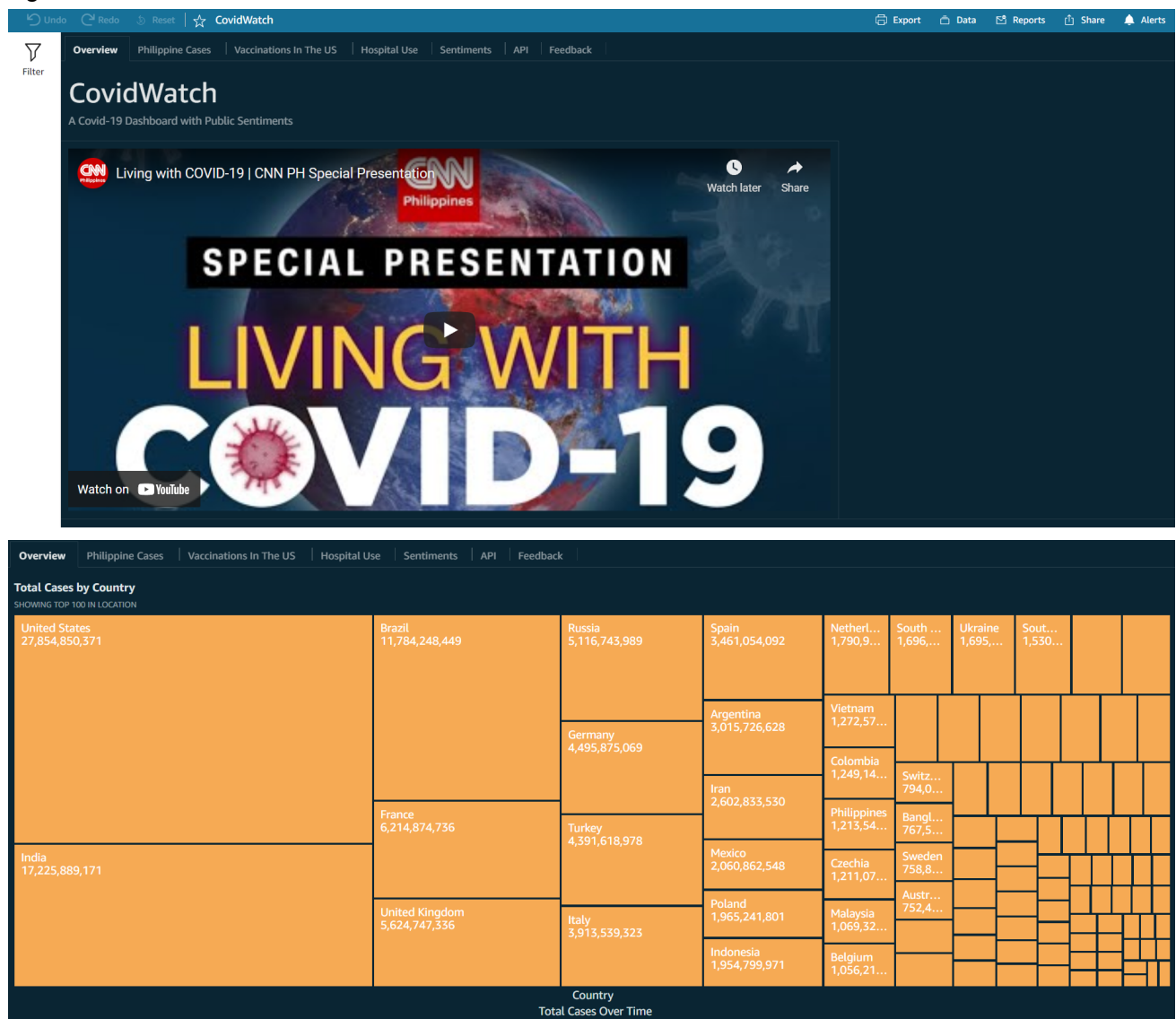
Figure 7. Overview of the Amazon QuickSight workflow



1. Overview

The overview section shows the latest news regarding the Covid-19 pandemic and the summary of Covid-19 cases per country. This will aid the decision-makers in their target country. Snippets of the Overview section of the dashboard are shown in Figure 8.

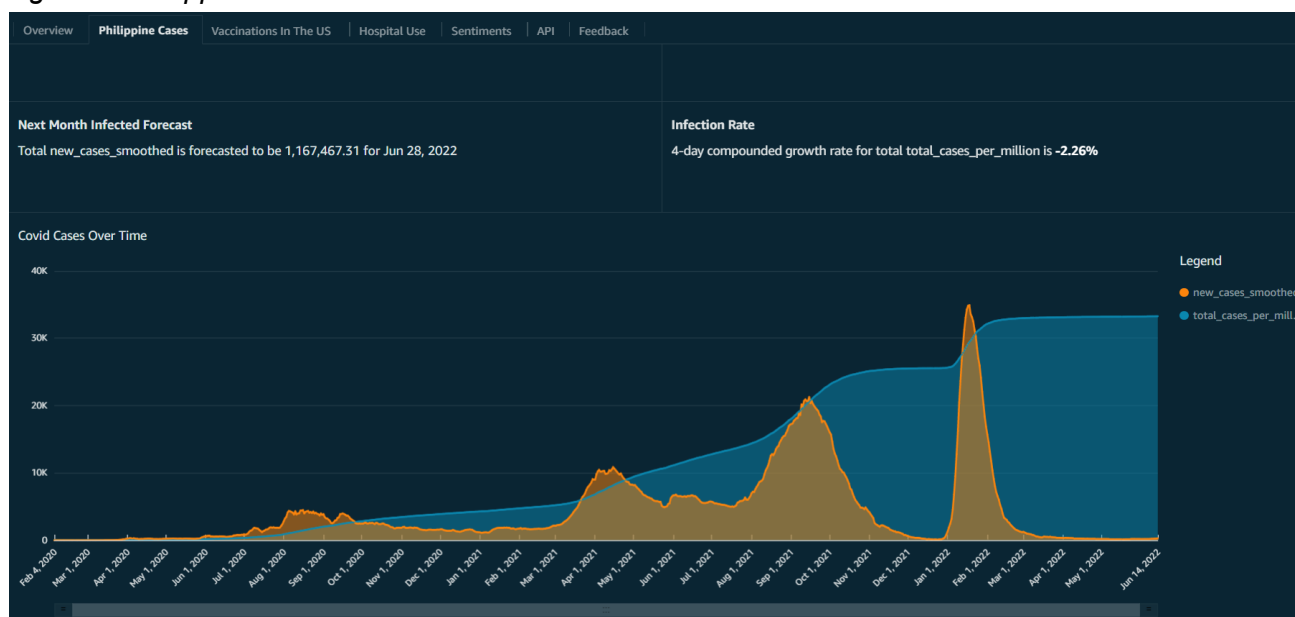
Figure 8. Overview section of the dashboard



2. Philippine Cases

To get a deeper understanding of the cases in a particular country, the Philippines, the team developed a section that shows the temporal aspect of the COVID-19. It shows the comparison of the new cases against the total cases. It also briefly summarizes the infection rate and the next month's infection forecast using the AWS machine learning. Snippets of the Philippine Cases section of the dashboard is shown in Figure 9.

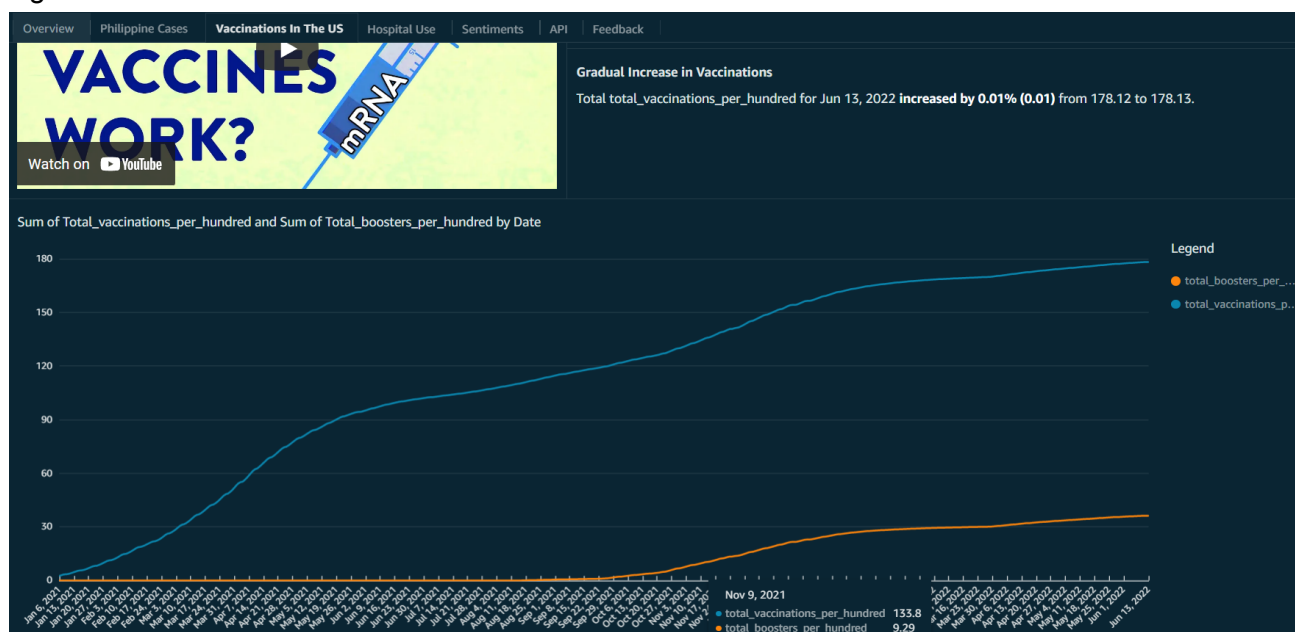
Figure 9. Philippine cases section of the dashboard



3. Vaccination in The US

This section provides a short video on how vaccines work to influence the public on its effect. It also shows the total vaccination and booster taken in the US. Snippets of the Vaccination in the US section of the dashboard is shown in Figure 10.

Figure 10. Vaccination in the US section of the dashboard

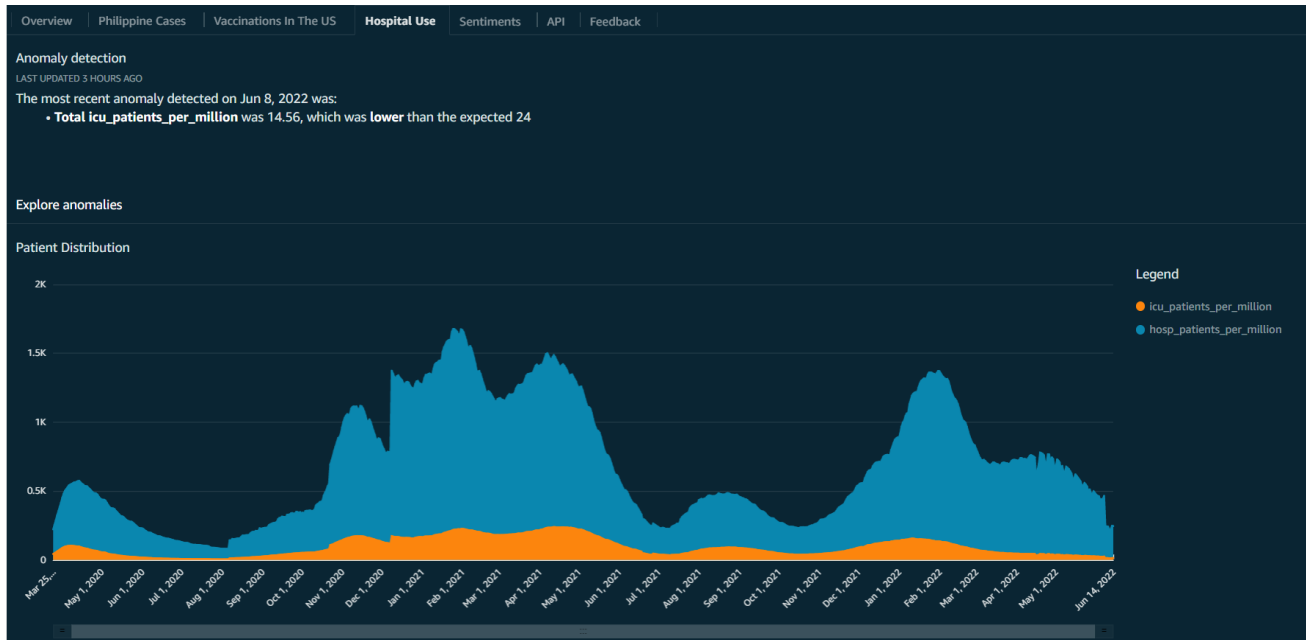


4. Hospital Use

It can also display hospital use data such as the number of total hospital patients over time and the number of ICU patients over time as well. Including this with the total capacity of hospitals will allow for policy makers to enact policies that can help alleviate the pressures on healthcare providers and services during dire times.

Anomaly detection can be seen here which is also provided by AWS. This can be used to see if there is an increase in the number of patients due to surges or the potency of symptoms such as that of a new variant. Snippets of the Hospital Use section of the dashboard is shown in Figure 11.

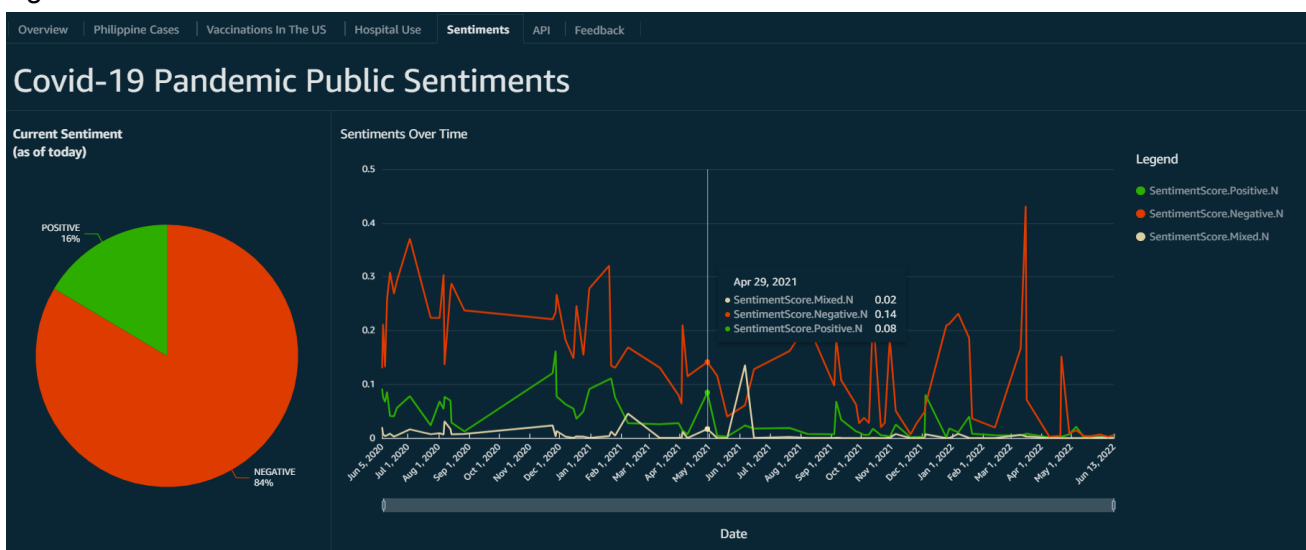
Figure 11. Vaccination in the US section of the dashboard



5. Public Sentiments

On the other side of the coin, using unconventional data sources such as tweets about covid, we can see the current sentiment of the public based on their tweets. We can also see it over time through the use of the line graph. The graph shows positive and negative sentiments. It can also show mixed and neutral scores. Snippets of the Public Sentiments section of the dashboard is shown in Figure 12.

Figure 12. Public Sentiments section of the dashboard



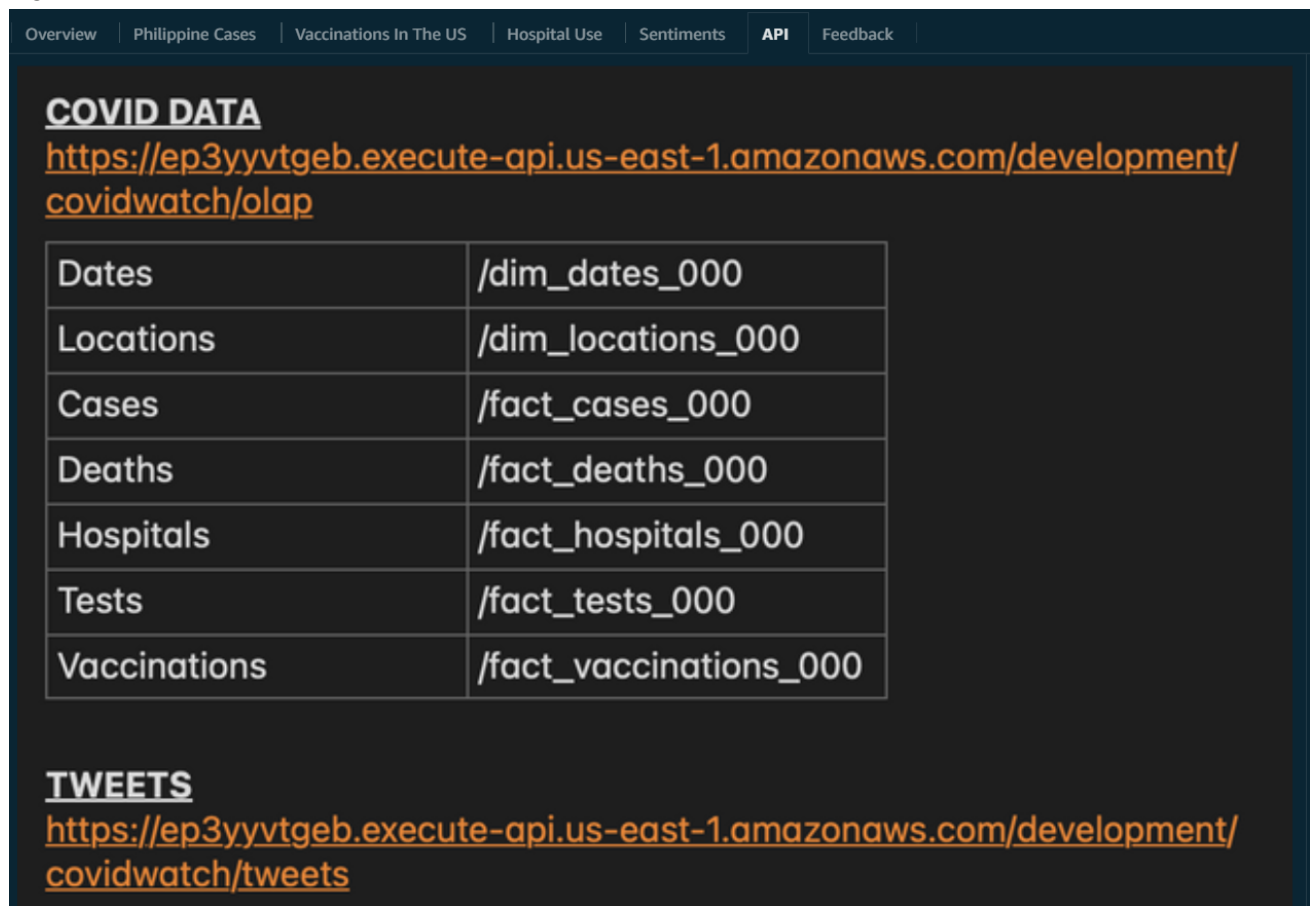


Here we can see the tweets that have the most positive and negative sentiments, who's size is dependent on the amount of engagement it received. These can be a more granular look at the public sentiment if there are any direct insights that need to be included in policy.

6. API Links

We also included the API links so that those who view the dashboard can access the data used behind it. Adding the string at the end of the URL selects different tables from the OLAP data. Snippets of the Public Sentiments section of the dashboard is shown in Figure 13.

Figure 13. API Links section of the dashboard



7. Feedback

For CovidWatch continuous improvement, the team created a feedback form to get the sentiments of the users that could possibly improve the dashboard in the future. Snippets of the Feedback section of the dashboard is shown in Figure 14.

Figure 14. Feedback section of the dashboard

Overview

Philippine Cases

Vaccinations In The US

Hospital Use

Sentiments



API

Feedback

Dashboard feedback

Thank you for using CovidWatch!

We want to hear your feedback so we can keep improving our logistics and content. Please fill this quick survey and let us know your thoughts (your answers will be anonymous).

 **pgplarosa@gmail.com** (not shared) [Switch account](#) 

*** Required**

How satisfied were you with CovidWatch? *

1

2

3

4

5

Not very

☐

☐

☐

☐

☐

Very much

How relevant and helpful do you think it was for your job? *

1

2

3

4

5

Not very

☐

☐

☐

☐

☐

Very much

VI. Conclusion

Vaccine hesitancy remains a significant hurdle to achieving herd immunity in many countries despite the rising number of daily COVID-19 cases worldwide and the proven efficacy of vaccines. To enable policymakers and health authorities to determine necessary actions in combating this public reluctance, it would be crucial for them to be aware of people's sentiments towards vaccines in addition to relevant COVID infection statistics.

This need for easily-accessible sentiment information is what inspired us to create CovidWatch. CovidWatch is an automated end-to-end database solution that extracts reliable data on COVID statistics and people's sentiments on a daily basis and provides policymakers and health analysts the means to derive meaningful and actionable insights through a dashboard and an API that allows access to raw data contained in its databases.

CovidWatch's architecture was designed with special attention to the system's scalability, evolvability and simplicity. CovidWatch meets these requirements through utilizing managed database solutions provided by AWS and creating the necessary components. CovidWatch contains three databases, each created for a different purpose. OLTP database allows CovidWatch to accommodate updates from various sources as they arrive. Meanwhile, the OLAP database implemented through Redshift enables faster querying with its columnar format and allows dimensional analysis with its location and time dimension tables that provide policymakers and stakeholders suitable statistics that apply to their jurisdiction and use case. Lastly, the NoSQL Database - the schemaless design of the NoSQL database enables preservation of all tweet information and provides maximum flexibility by allowing use cases beyond the sentiment analysis employed in this project.

For data storage, CovidWatch utilizes a data lake implemented through Amazon S3 with two zones that cater to different needs. The gold zone enables business analysts to easily extract meaningful insights from easy-to-use purpose-built data. Meanwhile, the landing zone preserves unaltered information for use by data scientists in advanced analytics and machine learning work.

ETL jobs created using Amazon Airflow perform the automated daily update of all the databases. For COVID data, DAGs download the data, perform ETL, and store the processed data on RDS, Redshift, and the S3 data lake . Meanwhile, another set of DAGs scrape tweets, perform sentiment analysis using Comprehend, and store results on both the S3 data lake and DynamoDB.

Finally, to bring value to its users, a Quicksight dashboard which updates automatically with the data is provided to analysts to aid in decision making. This dashboard not only provides valuable COVID statistics and sentiment data in just a few clicks, but also allows for limited machine learning capability. An API Access is also provided to the general public through a REST API, enabling them to take advantage of the raw data for any use case they desire.

Given these functionalities, CovidWatch demonstrates how the use of data engineering tools could assist with helping solve some of society's biggest problems by enabling the improvement of vaccine perception and alignment of policy with public opinion through providing accessibility of information surrounding covid and vaccines.

VII. Recommendation

Despite its strengths, many improvements can still be made to CovidWatch. One such possible refinement is its data source. As stated previously, the OLTP database currently draws data from the OWID Covid dataset through a daily download of the mega file which contains the historical data consolidated by OWID from many different sources. However, since future implementations are envisioned to receive daily updates from sources like hospitals, testing centers, and government monitoring agencies, the DAG implementation would need another ETL pipeline in the ingestion process to account for transforming the diverse data into a unified format. This new pipeline would be highly dependent on individual formats from the data received from the various sources.

For the OLAP database, the availability of more granular location data would require updating the location table which currently operates on a country level. This includes adding columns like hospital, city/municipality, state, region, etc. This more granular dimension would also require a different aggregation method which would have to be reflected on the DAG.

As for the sentiment data, Tweet data is currently retrieved using the hashtag #covid19usa which might not represent the sentiments of the whole country. Inclusion of other hashtags and other countries could improve the representation of the data and remove any bias that may occur. In addition, inclusion of other possible sources like GDELT, Facebook posts, and Instagram posts would greatly improve the representativeness of the sentiments it reflects.

VIII. References

- Cordero Jr, D. A. (2022). A more contextualized approach: Addressing COVID-19 vaccine hesitancy in the Philippines. *Infection & Chemotherapy*, 54(1), 178.
- Hannah Ritchie, Edouard Mathieu, Lucas Rod  s-Guirao, Cameron Appel, Charlie Giattino, Esteban Ortiz-Ospina, Joe Hasell, Bobbie Macdonald, Diana Beltekian and Max Roser (2020) - "Coronavirus Pandemic (COVID-19)". Published online at OurWorldInData.org. Retrieved from: '<https://ourworldindata.org/coronavirus>'
- Meng, M. D., & Olsen, M. C. (2021). Market segmentation strategies can be used to overcome COVID-19 vaccine hesitancy and other health crises. *Journal of Consumer Affairs*.
- Villavicencio C, Macrohon JJ, Inbaraj XA, Jeng J-H, Hsieh J-G. Twitter Sentiment Analysis towards COVID-19 Vaccines in the Philippines Using Na  ve Bayes. *Information*. 2021; 12(5):204. <https://doi.org/10.3390/info12050204>

IX. Appendix

Appendix A

Variable	Description
aged_65_older	Share of the population that is 65 years and older, most recent year available
aged_70_older	Share of the population that is 70 years and older in 2015
cardiovasc_death_rate	Death rate from cardiovascular disease in 2017 (annual number of deaths per 100,000 people)
continent	Continent of the geographical location
date	Date of observation
diabetes_prevalence	Diabetes prevalence (% of population aged 20 to 79) in 2017
excess_mortality	Percentage difference between the reported number of weekly or monthly deaths in 2020–2021 and the projected number of deaths for the same period based on previous years.
excess_mortality_cumulative	Percentage difference between the cumulative number of deaths since 1 January 2020 and the cumulative projected deaths for the same period based on previous years. For more information, see https://github.com/owid/covid-19-data/tree/master/public/data/excess_mortality
excess_mortality_cumulative_absolute	Cumulative difference between the reported number of deaths since 1 January 2020 and the projected number of deaths for the same period based on previous years. For more information, see https://github.com/owid/covid-19-data/tree/master/public/data/excess_mortality
extreme_poverty	Share of the population living in extreme poverty, most recent year available since 2010
female_smokers	Share of women who smoke, most recent year available
gdp_per_capita	Gross domestic product at purchasing power parity (constant 2011 international dollars), most recent year available
handwashing_facilities	Share of the population with basic handwashing facilities on premises, most recent year available
hosp_patients	Number of COVID-19 patients in hospital on a given day
hospital_beds_per_thous_and	Hospital beds per 1,000 people, most recent year available since 2010
human_development_index	A composite index measuring average achievement in three basic dimensions of human development—a long and healthy life, knowledge and a decent standard of living. Values for 2019, imported from http://hdr.undp.org/en/indicators/137506

icu_patients	Number of COVID-19 patients in intensive care units (ICUs) on a given day
iso_code	ISO 3166-1 alpha-3 – three-letter country codes
life_expectancy	Life expectancy at birth in 2019
location	Geographical location
male_smokers	Share of men who smoke, most recent year available
median_age	Median age of the population, UN projection for 2020
new_cases	New confirmed cases of COVID-19. Counts can include probable cases, where reported. In rare cases where our source reports a negative daily change due to a data correction, we set this metric to NA.
new_cases_smoothed	New confirmed cases of COVID-19 (7-day smoothed). Counts can include probable cases, where reported.
new_deaths	New deaths attributed to COVID-19. Counts can include probable deaths, where reported. In rare cases where our source reports a negative daily change due to a data correction, we set this metric to NA.
new_deaths_smoothed	New deaths attributed to COVID-19 (7-day smoothed). Counts can include probable deaths, where reported.
new_people_vaccinated_smoothed	Daily number of people receiving their first vaccine dose (7-day smoothed)
new_tests	New tests for COVID-19 (only calculated for consecutive days)
new_tests_smoothed	New tests for COVID-19 (7-day smoothed). For countries that don't report testing data on a daily basis, we assume that testing changed equally on a daily basis over any periods in which no data was reported. This produces a complete series of daily figures, which is then averaged over a rolling 7-day window
new_vaccinations	New COVID-19 vaccination doses administered (only calculated for consecutive days)
new_vaccinations_smoothed	New COVID-19 vaccination doses administered (7-day smoothed). For countries that don't report vaccination data on a daily basis, we assume that vaccination changed equally on a daily basis over any periods in which no data was reported. This produces a complete series of daily figures, which is then averaged over a rolling 7-day window
people_fully_vaccinated	Total number of people who received all doses prescribed by the initial vaccination protocol
people_vaccinated	Total number of people who received at least one vaccine dose
population	Population (latest available values)

population_density	Number of people divided by land area, measured in square kilometers, most recent year available
positive_rate	The share of COVID-19 tests that are positive, given as a rolling 7-day average (this is the inverse of tests_per_case)
tests_per_case	Tests conducted per new confirmed case of COVID-19, given as a rolling 7-day average (this is the inverse of positive_rate)
tests_units	Units used by the location to report its testing data
total_boosters	Total number of COVID-19 vaccination booster doses administered (doses administered beyond the number prescribed by the vaccination protocol)
total_cases	Total confirmed cases of COVID-19. Counts can include probable cases, where reported.
total_deaths	Total deaths attributed to COVID-19. Counts can include probable deaths, where reported.
total_tests	Total tests for COVID-19
total_vaccinations	Total number of COVID-19 vaccination doses administered
weekly_hosp_admissions	Number of COVID-19 patients newly admitted to hospitals in a given week (reporting date and the preceeding 6 days)
weekly_icu_admissions	Number of COVID-19 patients newly admitted to intensive care units (ICUs) in a given week (reporting date and the preceeding 6 days)