



Meet the Team

Bono
Jake
Hans
Nico
Neil

Sprint 2 Final Project

Group 1 - One Direction

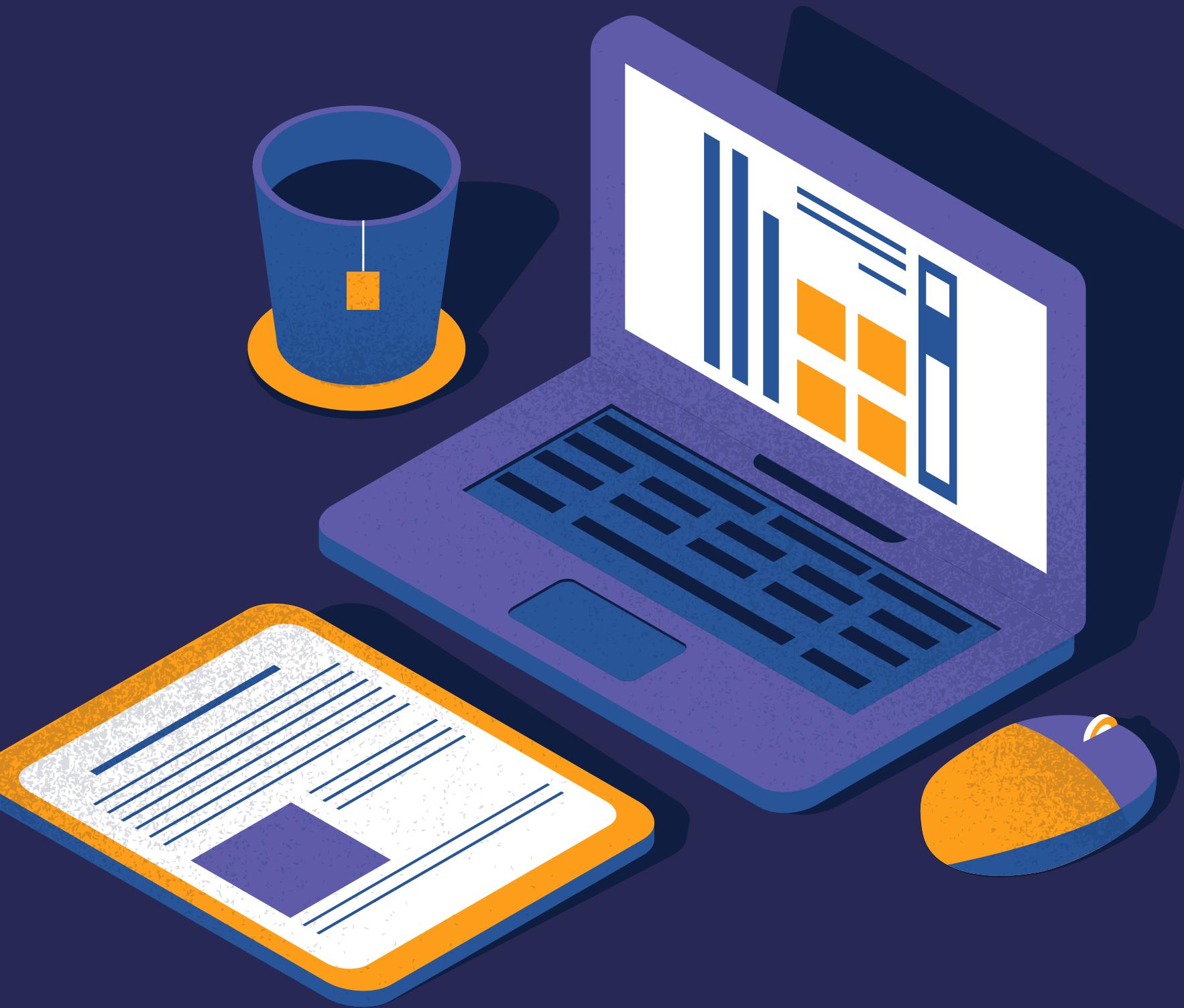


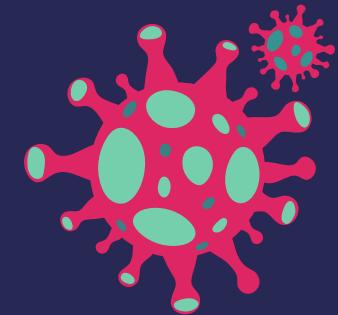


Agenda

- 1 Context of the Problem
- 2 Proposed DE Solution
- 3 Description of Features
Data Structure, Sources
- 4 DE Design & Results
ETL, Architecture
- 5 Discussion & Documentation

Context of the Problem





3.6m infected, 60k dead



2.4 million at risk

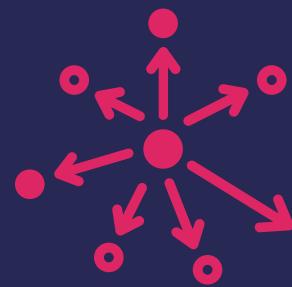


**Need for data-driven
decision-making**

Source: World Health Organization, Philippines ; United Nations Office for Coordinated Humanitarian Affairs



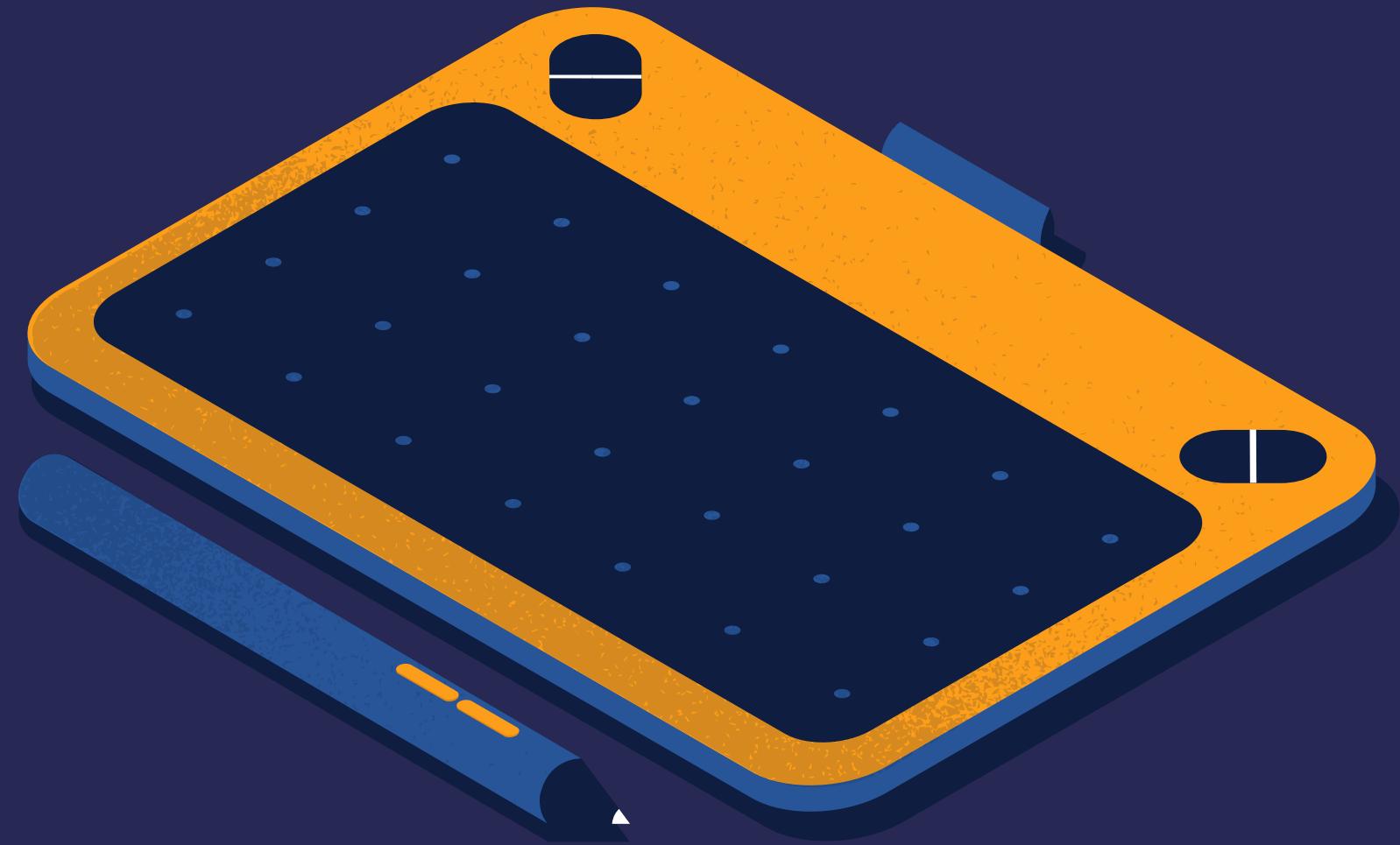
Inaccessible data formats



Raw, unprocessed, and
decentralized



Slow, uncertain decision-
making



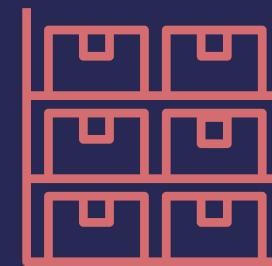
Proposed DE Solution



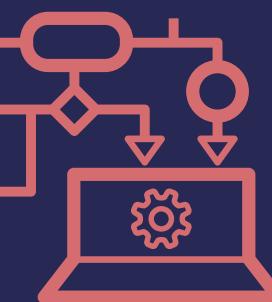
Centralized data pipeline



Automates data collection and parsing



Accessible data mart



Normalized data for reporting and model-building



Evidence for ML developed solutions



XGBoost to quantify COVID-19 risk for counties

Source: Mehta, Julaiti, Griffin & Kumara (2020)



Ensemble models and multi-objective optimization to forecast disease prevalence

Source: Tiwari et al. (2021) ; Pan et al. (2021)

Description of Features





Data Sources

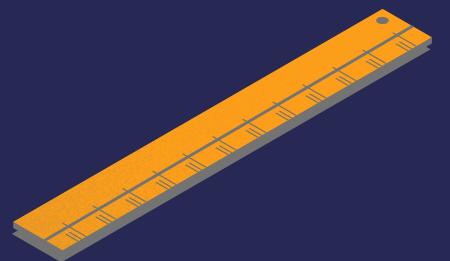


Table Schema



Application



Data Sources



External Data



Web Scraping



APIs



External Data

Source: [DOH covid19 tracker](#)

COVID-19 Tracker
PHILIPPINES

[f /OfficialDOHgov](#) [@DOHgovph](#)

[Download COVID-19 DOH Data Drop](#)

Overview

Filter Epidemiology data by Region: (All) by Province / HUC / ICC: (All)

As of June 10, 2022

HUC - "Highly Urbanized City"
ICC - "Independent Component City"

Nationwide Cases Data

Total Cases 3,692,617 +287 added on 06/10	Active Cases 2,697	Recovered 3,629,459	Died 60,461	View Detailed Case Information
--	------------------------------	-------------------------------	-----------------------	--

Confirmed cases are those that tested RT-PCR positive by a DOH-RITM certified lab.

Weekly Cases by Date of Onset of Illness

For 63.0% or 2,324,935 of cases where date of onset of illness is unreported, date of specimen collection was used as proxy.

Weekly Daily

Cases Recoveries Deaths

We urge caution when interpreting data during the highlighted period below, which may be incomplete because of delays in reporting.

- 4-Week Moving Average

Feb 1, 20 Aug 1, 20 Nov 1, 20 Feb 1, 21 May 1, 21 Aug 1, 21 Nov 1, 21 Feb 1, 22 May 1, 22

Note: There are still 720113 cases with unreported date of onset of illness and date of specimen collection.

Features:

- Regional COVID Case information
- Reported staff counts
- Reported IPC standards implementation

CaseCode	Age	AgeGroup	Sex	DateSpecimen	DateResultRelease	DateRepConf	DateDied	DateRecover	RemovalType	A
C404174	38	35 to 39	FEMALE		2020-01-30	2020-01-30			RECOVERED	
C462688	44	40 to 44	MALE		2020-01-30	2020-02-03	2020-02-01		DIED	
C387710	60	60 to 64	FEMALE	2020-01-23	2020-01-30	2020-02-05		2020-01-31	RECOVERED	
C377460	49	45 to 49	MALE			2020-03-06			RECOVERED	
C498051	63	60 to 64	MALE	2020-03-05		2020-03-06	2020-03-11		DIED	
C130591	58	55 to 59	FEMALE	2020-03-06	2020-03-07	2020-03-12			DIED	
C440075	33	30 to 34	MALE	2020-03-06	2020-03-08	2020-03-08		2020-04-05	RECOVERED	
C202135	57	55 to 59	MALE	2020-03-06	2020-03-08	2020-03-08		2020-03-23	RECOVERED	
C178743	39	35 to 39	MALE	2020-03-06	2020-03-08	2020-03-08		2020-03-21	RECOVERED	
C557002	86	80+	MALE	2020-03-06	2020-03-08	2020-03-08	2020-03-14		DIED	
C787672	70	70 to 74	MALE	2020-03-07	2020-03-09	2020-03-09		2020-03-27	RECOVERED	
C325527	59	55 to 59	FEMALE				2020-03-09		RECOVERED	
C777589	31	30 to 34	FEMALE				2020-03-09		RECOVERED	
C557823	41	40 to 44	MALE			2020-03-09		2020-03-24	RECOVERED	
C348794	70	70 to 74	MALE	2020-03-06	2020-03-09	2020-03-09		2020-03-19	RECOVERED	



External Data

Source: Philippine Statistics Authority

- Population & annual growth rate
- Urbanized City Status
- Region, Province, and City/Municipality
- Poverty Incidence Among Families

A. POPULATION AND ANNUAL GROWTH RATE FOR THE PHILIPPINES AND ITS REGIONS, PROVINCES, AND HIGHLY URBANIZED CITIES BASED ON THE 2000, 2010, 2015, AND 2020 CENSUSES								
REGION, PROVINCE, AND HIGHLY URBANIZED CITY	TOTAL POPULATION				POPULATION GROWTH RATE (in percent)			
	01-May-00	01-May-10	01-Aug-15	01-May-20	2000-2010	2010-2015	2015-2020	2010-2020
PHILIPPINES	76,506,928 ^a	92,337,852 ^b	100,981,437 ^c	109,035,343 ^d	1.90	1.72	1.63	1.67
NATIONAL CAPITAL REGION (NCR)	9,932,560	11,855,975	12,877,253	13,484,462	1.78	1.58	0.97	1.29
CITY OF MANILA	1,581,082	1,652,171	1,780,148	1,846,513	0.44	1.43	0.77	1.12
CITY OF MANDALUYONG	278,474	328,699	386,276	425,758	1.67	3.12	2.07	2.62
CITY OF MARIKINA	391,170	424,150	450,741	456,059	0.81	1.16	0.25	0.73

REGION, PROVINCE, AND CITY/MUNICIPALITY	TOTAL POPULATION				POPULATION GROWTH RATE (in percent)		
	01-May-00	01-May-10	01-Aug-15	01-May-20	2000-2010	2010-2015	2015-2020
REGION VIII (EASTERN VISAYAS)	3,610,355	4,101,322	4,440,150	4,547,150	1.28	1.52	
BILIRAN	140,274	161,760	171,612	179,312	1.43	1.13	
ALMERIA	13,854	16,495	16,951	17,954	1.76	0.52	
BILIRAN	13,817	16,183	16,882	17,662	1.59	0.81	
CABUCGAYAN	17,691	19,621	20,788	21,542	1.04	1.11	
CAIBIRAN	19,606	21,473	22,524	24,167	0.91	0.91	
CULABA	11,506	12,252	12,325	12,972	0.63	0.11	
KAWAYAN	17,507	20,238	20,291	20,455	1.46	0.05	
MARIBUAO	2,242	2,389	2,450	2,472	0.14	0.07	



External Data

Source: Our World In Data

Features:

- Dates
- Vaccine Used
- Source of Information
- Total Vaccinations
- People Vaccinated
- People Fully Vaccinated
- Total Booster Shots

location	date	vaccine	source_url	total_vacc	people_va	people_fu	total_boos
Philippine	14/10/2021	Johnson& https://news.abs-cbn.com/spotli	51482063		23981240		
Philippine	17/10/2021	Johnson& https://news.abs-cbn.com/spotli	52303905		24307903		
Philippine	18/10/2021	Johnson& https://news.abs-cbn.com/spotli	52783354		24498753		
Philippine	19/10/2021	Johnson& https://news.abs-cbn.com/spotli	53315069		24694717		
Philippine	20/10/2021	Johnson& https://news.abs-cbn.com/spotli	53838248		24876889		
Philippine	21/10/2021	Johnson& https://covid19.who.int	54444161	32954936	25101222		
Philippine	24/10/2021	Johnson& https://news.abs-cbn.com/spotli	55715693		25711980		
Philippine	25/10/2021	Johnson& https://news.abs-cbn.com/spotli	56254529		25955669		
Philippine	26/10/2021	Johnson& https://news.abs-cbn.com/spotli	56774753		26180669		
Philippine	27/10/2021	Johnson& https://news.abs-cbn.com/spotli	57494154		26479028		
Philippine	28/10/2021	Johnson& https://news.abs-cbn.com/spotli	58212187		26803677		
Philippine	31/10/2021	Johnson& https://news.abs-cbn.com/spotli	59316764		27360873		
Philippine	01/11/2021	Johnson& https://news.abs-cbn.com/spotli	59473662		27442969		
Philippine	02/11/2021	Johnson& https://news.abs-cbn.com/spotli	60406424		27749809		
Philippine	03/11/2021	Johnson& https://news.abs-cbn.com/spotli	61354945		28198294		
Philippine	04/11/2021	Johnson& https://news.abs-cbn.com/spotli	62474334				
Philippine	07/11/2021	Johnson& https://news.abs-cbn.com/spotli	64195936				
Philippine	08/11/2021	Johnson& https://news.abs-cbn.com/spotli	64947366				
Philippine	09/11/2021	Johnson& https://news.abs-cbn.com/spotli	65764376				
Philippine	10/11/2021	Johnson& https://news.abs-cbn.com/spotli	66816976				
Philippine	11/11/2021	Johnson& https://covid19.who.int	67716205	40517967	30804594		
Philippine	14/11/2021	Johnson& https://news.abs-cbn.com/spotli	69713994				
Philippine	15/11/2021	Johnson& https://news.abs-cbn.com/spotli	70677771				
Philippine	16/11/2021	Johnson& https://news.abs-cbn.com/spotli	71680132				
Philippine	17/11/2021	Johnson& https://news.abs-cbn.com/spotli	72763442				
Philippine	18/11/2021	Johnson& https://news.abs-cbn.com/spotli	73917573		32993083		
Philippine	21/11/2021	Johnson& https://news.abs-cbn.com/spotli	75600808		33579181		
Philippine	28/11/2021	Johnson& https://news.abs-cbn.com/spotli	81296947		35678774	188084	
Philippine	30/11/2021	Johnson& https://news.abs-cbn.com/spotli	86421420		36365357	313836	
Philippine	01/12/2021	Johnson& https://news.abs-cbn.com/spotli	89070292		36869419	389451	



Web Scraping

Scraping COVID disinformation

- Factraker
- Rappler.com

Features:

- Description
- Date and time posted
- URL source

	post_id	text	post_text	shared_text	original_text	time	timestamp	image	image_lowquality	images	...	reaction_count	with	page_id
32	5293592450732983	Marcos Gold, totoo raw ayon sa BSP? \n\nHINDI T...	Marcos Gold, totoo raw ayon sa BSP?			2022-05-05 14:43:02	1651732982	None	https://scontent.fmn16-1.fna.fbcdn.net/v/t39.3...	[None]	...	None	None	1624618274297104
82	5268013669957528	Diktador na si Marcos, nag-imbak daw ng ginto ...	Diktador na si Marcos, nag-imbak daw ng ginto ...			2022-04-26 07:00:04	1650927604	None	https://scontent.fmn16-2.fna.fbcdn.net/v/t39.3...	[None, None]	...	None	None	1624618274297104
110	5255752397850322	BSP Governor Diokno, pinatunayang totoo ang Ma...	BSP Governor Diokno, pinatunayang totoo ang Ma...			2022-04-21 17:00:04	1650531604	None	https://scontent.fmn16-1.fna.fbcdn.net/v/t39.3...	[None]	...	None	None	1624618274297104
135	5250328508392711	Leni, tumakbo lang para sa pagkapangulo para k...	Leni, tumakbo lang para sa pagkapangulo para k...			2022-04-19 16:00:02	1650355202	None	https://scontent.fmn16-2.fna.fbcdn.net/v/t39.3...	[None]	...	None	None	1624618274297104



APIs

COVID-19 statistics based on public data by John Hopkins CSSE

	date	confirmed	deaths	recovered	confirmed_diff	deaths_diff	recovered_diff	last_update	active	active_diff	fatality_rate
0	2020-04-12	4648	297	197	220	50	40	2020-04-12 23:17:00	4154	130	0.0639
1	2020-04-13	4932	315	242	284	18	45	2020-04-13 23:07:34	4375	221	0.0639
2	2020-04-14	5223	335	295	291	20	53	2020-04-14 23:33:12	4593	218	0.0641
3	2020-04-15	5453	349	353	230	14	58	2020-04-15 22:56:32	4751	158	0.0640
4	2020-04-16	5660	362	435	207	13	82	2020-04-16 23:30:31	4863	112	0.0640
5	2020-04-17	5878	387	487	218	25	52	2020-04-17 23:30:32	5004	141	0.0658
6	2020-04-18	6087	397	516	209	10	29	2020-04-18 22:32:28	5174	170	0.0652
7	2020-04-19	6259	409	572	172	12	56	2020-04-19 23:40:41	5278	104	0.0653
8	2020-04-20	6459	428	613	200	19	41	2020-04-20 23:36:27	5418	140	0.0663
9	2020-04-21	6599	437	654	140	9	41	2020-04-21 23:30:30	5508	90	0.0662
10	2020-04-22	6710	446	693	111	9	39	2020-04-22 23:30:32	5571	63	0.0665
11	2020-04-23	6981	462	722	271	16	29	2020-04-24 03:30:31	5797	226	0.0662
12	2020-04-24	7192	477	762	211	15	40	2020-04-25 06:30:33	5953	156	0.0663
13	2020-04-25	7294	494	792	102	17	30	2020-04-26 02:30:31	6008	55	0.0677
14	2020-04-26	7579	501	862	285	7	70	2020-04-27 02:30:33	6216	208	0.0661
15	2020-04-27	7777	511	932	198	10	70	2020-04-28 02:30:32	6334	118	0.0657
16	2020-04-28	7958	530	975	181	19	43	2020-04-29 02:32:29	6453	119	0.0666
17	2020-04-29	8212	558	1023	254	28	48	2020-04-30 02:32:27	6631	178	0.0679
18	2020-04-30	8488	568	1043	276	10	20	2020-05-01 02:32:28	6877	246	0.0669

Features:

1. Date
2. Confirmed Cases
3. Deaths
4. Recovered
5. Confirmed diff
6. Deaths diff
7. Recovered diff
8. Last update
9. Active cases
10. Active diff
11. Fatality rate

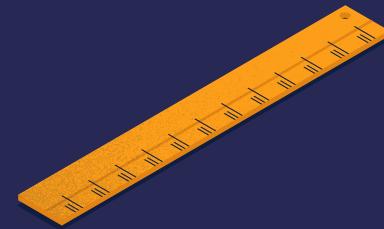
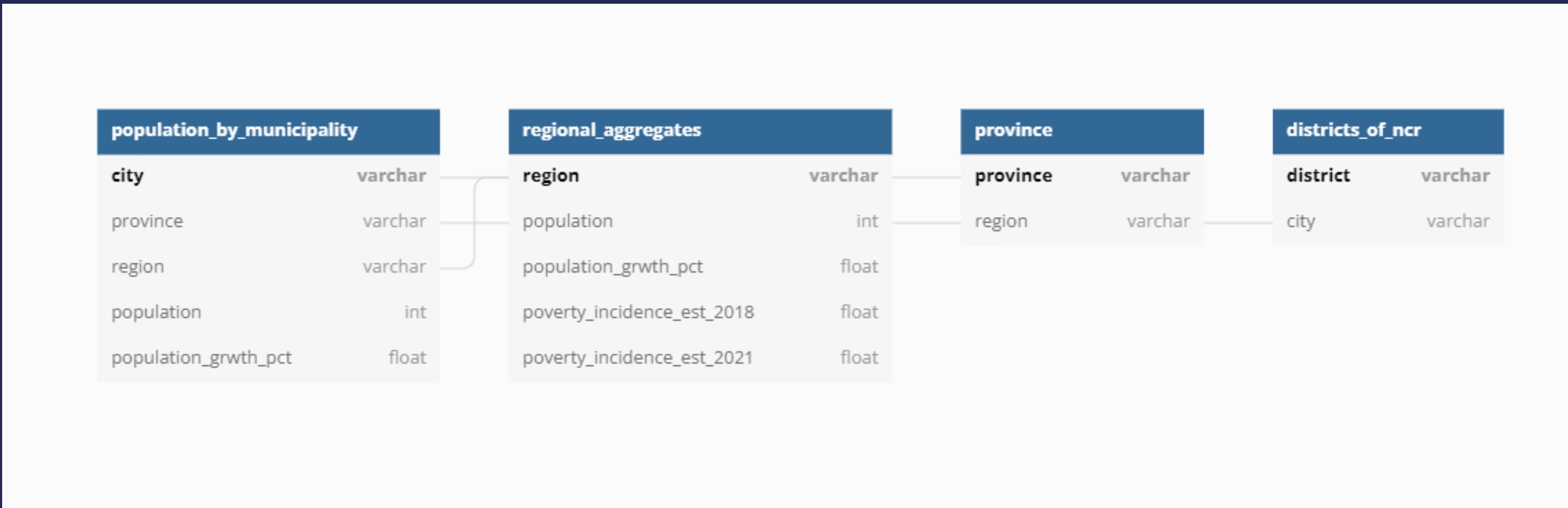
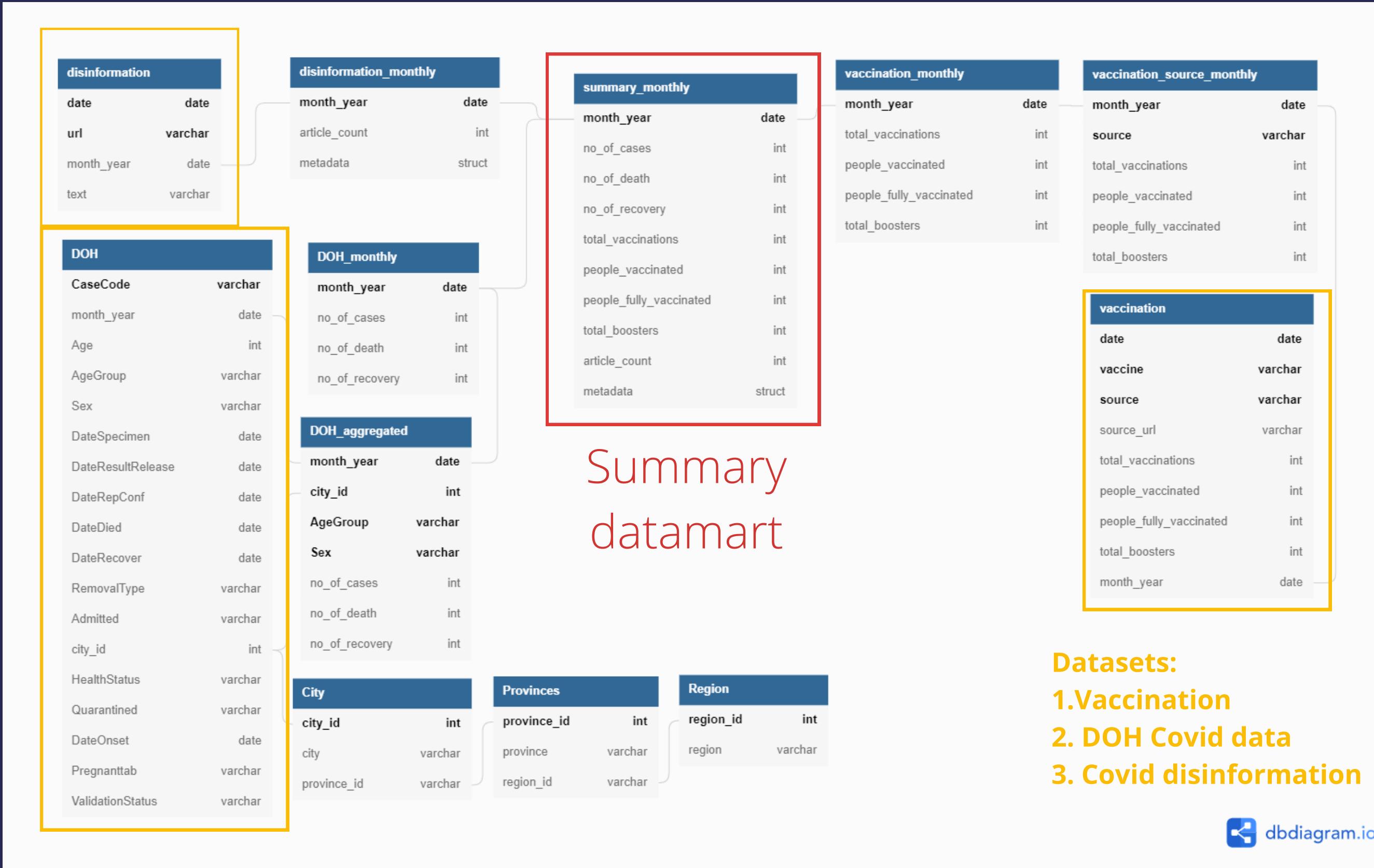
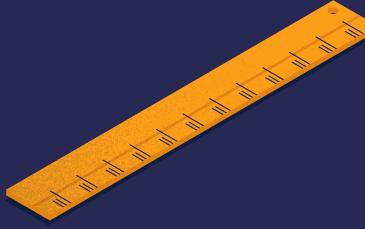


Table Schema - Socio-demo



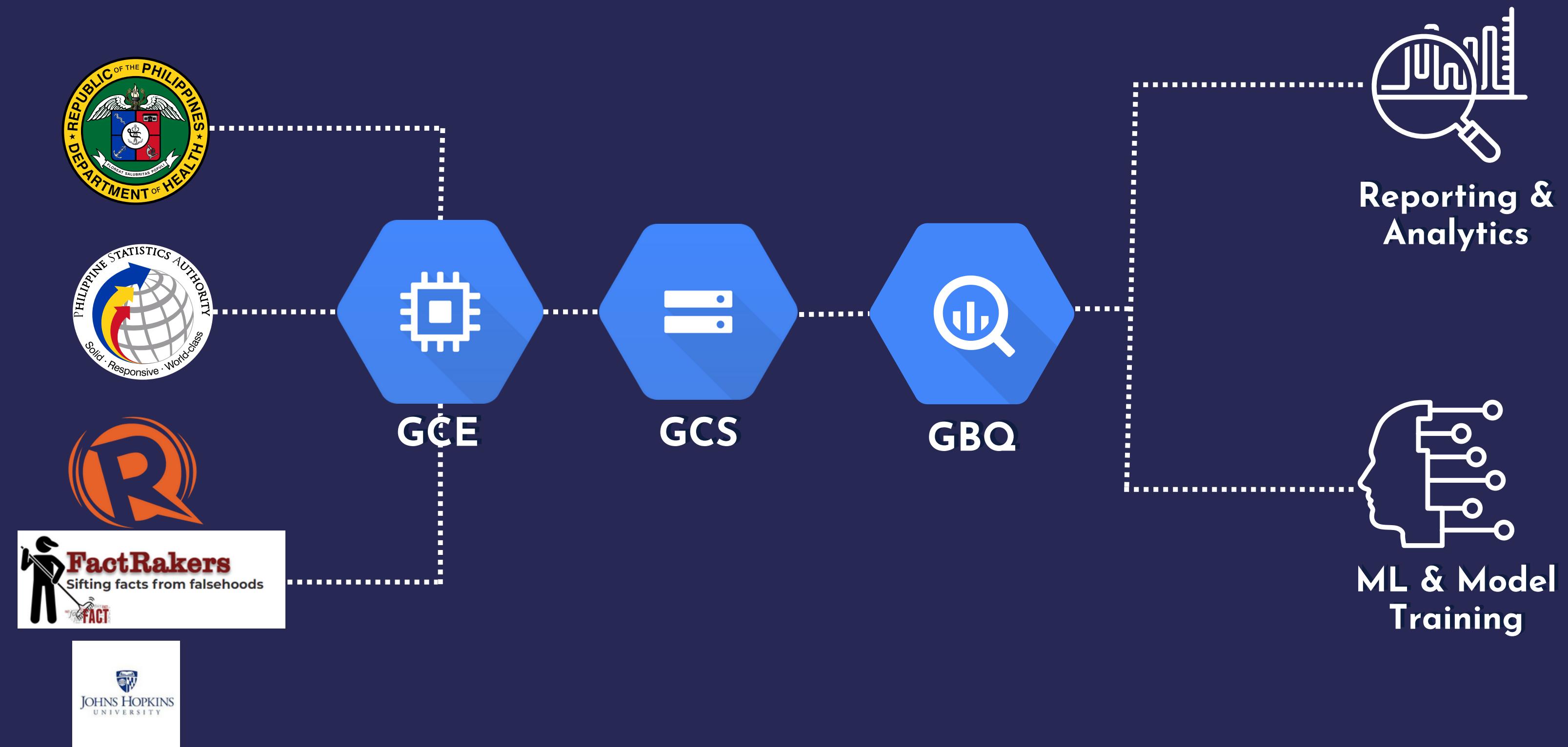
[https://dbdiagram.io
/d/62a2245592b33b4f51345b02](https://dbdiagram.io/d/62a2245592b33b4f51345b02)

Table Schema - health and disinformation





Application



DE Design & Results





Data Ingestion



Data Transformation



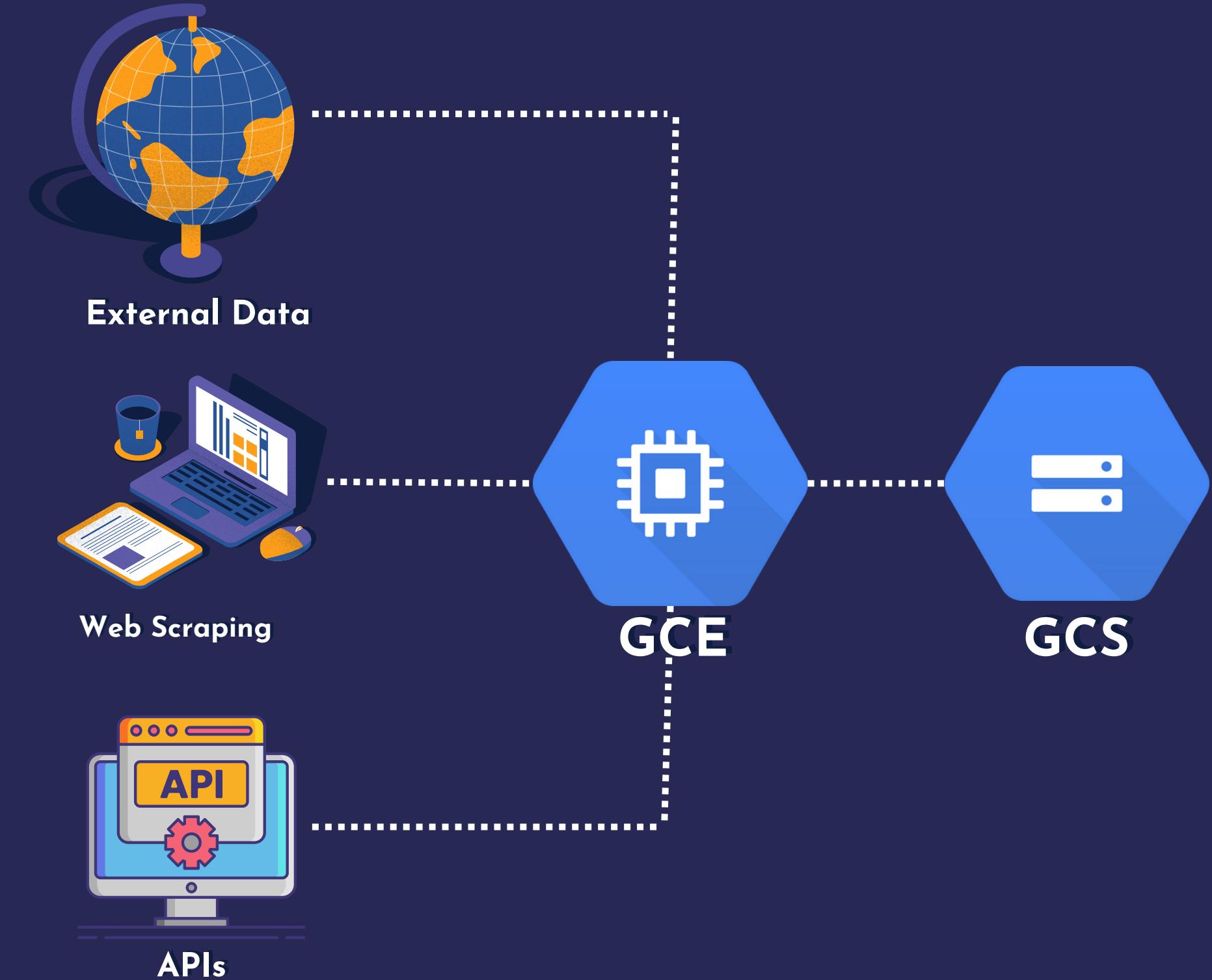
ERD



Loading Data

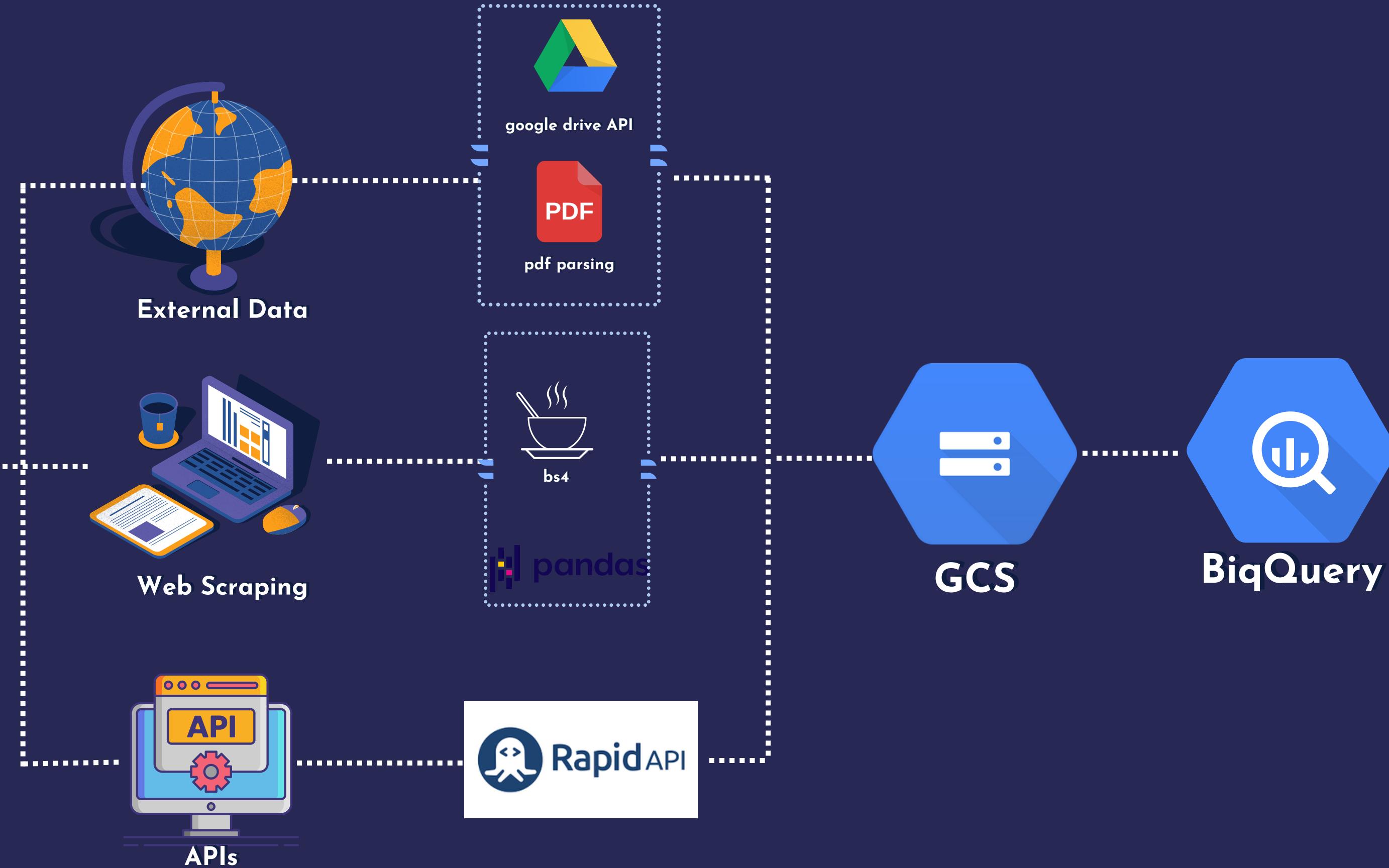


Data Ingestion





Data Flow





Data Transformation

Query results

SAVE RESULTS EXPLORE DATA

JOB INFORMATION RESULTS JSON EXECUTION DETAILS

⚠ Showing the first 500 columns for performance reasons. Consider modifying your query to show fewer columns. DISMISS

Row	i_3	aggregation_level	new_confirmed	new_deceased	cumulative_confirmed	cumulative_deceased	cumulative_tested	new_persons_vaccinated	cumulative_persons_vaccinate
1	0	211	2	2838901	51213	23690229	0	null	
2	0	24851	47	3441969	53519	25377563	null	null	
3	0	0	0	3	1	null	null	null	
4	0	0	0	0	0	null	null	null	
5	0	9868	0	2622917	38828	20371132	0	null	
6	0	2718	164	3639930	55094	26210790	0	null	
7	0	4352	135	2765631	42077	21435114	0	null	
8	0	173	0	3686475	60439	null	null	null	
9	0	18101	70	2402447	59796	25561255	null	null	



Data Transformation

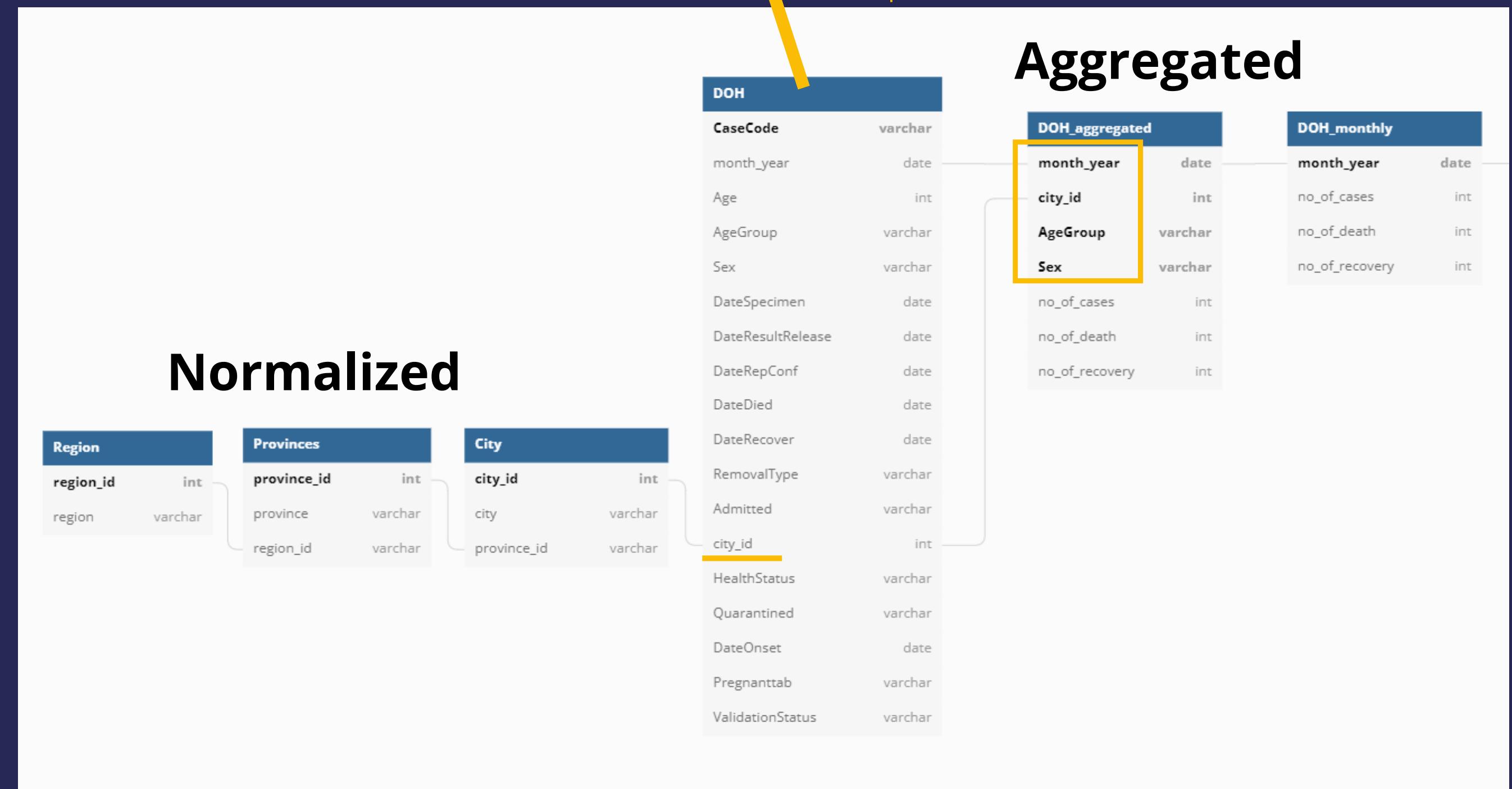
```
C: > Users > hxwoo > Desktop > sprint2-covid-project > socio-dem-data > add_urbanicity.py
1 import pandas as pd
2 import numpy as np
3
4 pop = pd.read_csv('population_by_municipality.csv')
5
6 pop['highly_urbanized'] = pop['province'].apply(lambda x: 1 if x == 'high' else 0)
7 pop['province'] = pop['province'].apply(lambda prov: 'NA' if prov == 'high' else prov)
8
9 pop.to_csv('population_by_municipality_final.csv', index=False)
```

A	B	C	D	E	F	G	H
city	province	region	populatio	populatio	highly_urbanized		
CITY OF MANILA	NA	NATIONAL	1846513	1.12	1		
CITY OF MANDALUYONG	NA	NATIONAL	425758	2.62	1		
CITY OF MARIKINA	NA	NATIONAL	456059	0.73	1		
CITY OF PASIG	NA	NATIONAL	803159	1.83	1		
QUEZON CITY	NA	NATIONAL	2960048	0.7	1		
CITY OF SAN JUAN	NA	NATIONAL	126347	0.4	1		
CITY OF CALOOCAN	NA	NATIONAL	1661584	1.1	1		
CITY OF MALABON	NA	NATIONAL	380522	0.74	1		
CITY OF NAVOTAS	NA	NATIONAL	247543	-0.06	1		
CITY OF VALENZUELA	NA	NATIONAL	714978	2.19	1		
CITY OF LAS PIÑAS	NA	NATIONAL	606293	0.93	1		
CITY OF MAKATI	NA	NATIONAL	629616	1.75	1		



ERD

DOH covid cases





Loading Data

Google Cloud Platform DEC1-Sprint2-COVID19DB ▾ Search Products, resources, docs (/) 1 ? ⋮ Profile

Explorer + ADD DATA ◀

vaccinati... ata X +

vaccination-data QUERY SHARE COPY SNAPSHOT DELETE EXPORT

SCHEMA DETAILS PREVIEW

Filter Enter property name or value

Field name	Type	Mode	Collation	Policy Tags ?	Description
date	DATE	NULLABLE			
vaccine	STRING	NULLABLE			
source_url	STRING	NULLABLE			
total_vaccinations	INTEGER	NULLABLE			
people_vaccinated	INTEGER	NULLABLE			
people_fully_vaccinated	INTEGER	NULLABLE			
total_boosters	INTEGER	NULLABLE			

EDIT SCHEMA VIEW ROW ACCESS POLICIES

Viewing pinned projects.

- dec1-sprint2-covid19db
 - covid19_data
 - vaccination-data
 - disinformation
 - factracker-results
 - rappler-results
 - doh_data
 - socdem_data
 - ncr_districts
 - population_by_municipality
 - province_region
 - regional_aggregates
- bigquery-public-data
- dab-c4-payruler
- dabc4-sprint2

Discussion & Documentation



Demo of Application

50 lines (34 sloc) | 3.67 KB

<>

COV1D-19 Philippine Data Pipeline

This project was done in accordance with the requirements of Sprint 2 of the Data Engineering Bootcamp - Cohort 1.

This group is composed of Bono, Jake, Hans, Nico, and Neil.

Description and Objective

These are the scripts used for a centralized data pipeline using Philippine COVID-19 data that is deployed in our GCP project for the second sprint for the ESK DEBC1.

The scripts are run once day at XX:XX:XX AM UTC+8

Data Sources: This list is the complete list of sources that the data is pulled from.

- Aggregated data sources:
 - Fact Rackers News: <https://www.factrakers.org/search-results?q=Covid>
 - Covid 19- Fact Checks Rappler: <https://www.rappler.com/topic/covid-19-fact-checks/>
 - COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University: <https://github.com/CSSEGISandData/COVID-19>
 - Our World In Data: https://raw.githubusercontent.com/owid/covid-19-data/master/public/data/vaccinations/country_data/Philippines.csv
- Non-aggregated data sources:
 - Republic of Philippines Department of Health: <https://doh.gov.ph/covid19tracker>
 - Philippine Statistics Authority: <https://psa.gov.ph/issip-stat-domain/demographic-and-social-statistics>
- The data from the Philippine Statistics Authority was manually edited into tidy data format using Google Sheets and Microsoft Excel. This approach was necessary due to the variance in formatting and the presence of merged cells.
- A short script, `add_urbanicity.py` was added to create a new data field `highly_urbanized` that identifies whether or not a particular city is classified as highly urbanized according to the Philippine Statistics Authority. This metric of urbanization is a relevant feature due to its relative prominence in predicting COVID-19 vulnerability and preparedness.

<https://github.com/hxwwong/sprint2-covid-project/blob/main/README.md>

Thank you for listening!

