A background image showing several COVID-19 test tubes. One tube in the foreground is clearly labeled "COVID-19" and "TEST".

# COVID IMPACT ANALYSIS

using Hadoop, Apache Hive & Excel

Group 3: Bryan Chica  
Mohammed Shodimu  
Phillip Navarro  
Suleima Gonzalez

CIS 4560-01  
Dr. Jongwook Woo  
December 3, 2025

# Agenda

- Introduction
- Dataset Overview
- Platform Specifications
- Architecture
- Workflow / Flowchart
- Tempo-Spatial Visualizations
- Analysis & Findings
- Hive Processing
- Github Repository
- Conclusion



# Introduction

This Big data project studies the global impact of COVID-19 by analyzing a large Google Public COVID-19 Open dataset. We implemented tempo-spatial analysis to understand how the virus affected people with regards to:

- *Different groups i.e age, gender, geography*
- *Health availability and expenditures*
- *Varying measures i.e. Vaccine doses*

Relevance: Using tools such as Hive, HDFS and Power BI we processed millions of records to identify meaningful patterns and problem areas in COVID's impact on healthcare, economic and social sectors worldwide.

# Dataset Specifications

## Dataset

- **Source:** Google Public Dataset
- **File name:** aggregated.csv
- **Compressed size:** 1.32 GB
- **Extracted size:** 20.9 GB
- **Records:** 3,000,599
- **Columns:** 708
- **Format:** .csv
- **Data Types:** STRING, FLOAT, INT, BOOLEAN, DATETIME
- **Link:** <https://storage.googleapis.com/covid19-open-data>

## Github

### Includes

- **Hive .hql scripts**
- **CSV outputs**
- **Hadoop commands**
- **Workflow documentation**
- **Presentation Slides**
- **Architecture diagrams**
- **Flowchart**
- **Link:** <https://github.com/chicagohg/Covid19-Data-Analysis>

```
-bash-4.2$  
-bash-4.2$ hostnamectl | grep "hostname"  
  Static hostname: bigdaiun0.sub03291929060.trainingvcn.oraclevcn.com  
-bash-4.2$  
-bash-4.2$  
-bash-4.2$ hostnamectl | grep "Operating System"  
  Operating System: Oracle Linux Server 7.9  
-bash-4.2$
```

## Host name & Operating System

- **Hostname:**  
bigdaiun0.sub03291929060.trainingvcn.oraclevcn.com
- **OS:** Oracle Linux Server 7.9

```
-bash-4.2$ lscpu
Architecture:          x86_64
CPU op-mode(s):        32-bit, 64-bit
Byte Order:            Little Endian
CPU(s):                6
On-line CPU(s) list:  0-5
Thread(s) per core:   2
Core(s) per socket:   3
Socket(s):             1
NUMA node(s):          1
Vendor ID:             AuthenticAMD
CPU family:            25
Model:                 1
Model name:            AMD EPYC 7J13 64-Core Processor
Stepping:               1
CPU MHz:               2445.406
```

## System Resources

- **CPU:** AMC EPYX 7J13 64-Core Processor @ 2.45 Ghz
- **Storage:** 120 GB
- **Memory Size:** 32 GB (8 GB swap)

```
-bash-4.2$ df -h ~
Filesystem      Size  Used Avail Use% Mounted on
/dev/sda3       120G  103G  18G  86% /
-bash-4.2$ free -h
              total        used         free      shared  buff/cache   available
Mem:           31G         26G        739M      507M        4.2G        4.1G
Swap:          8.0G        2.6G        5.4G
```



Free storage space

## Other Specifications

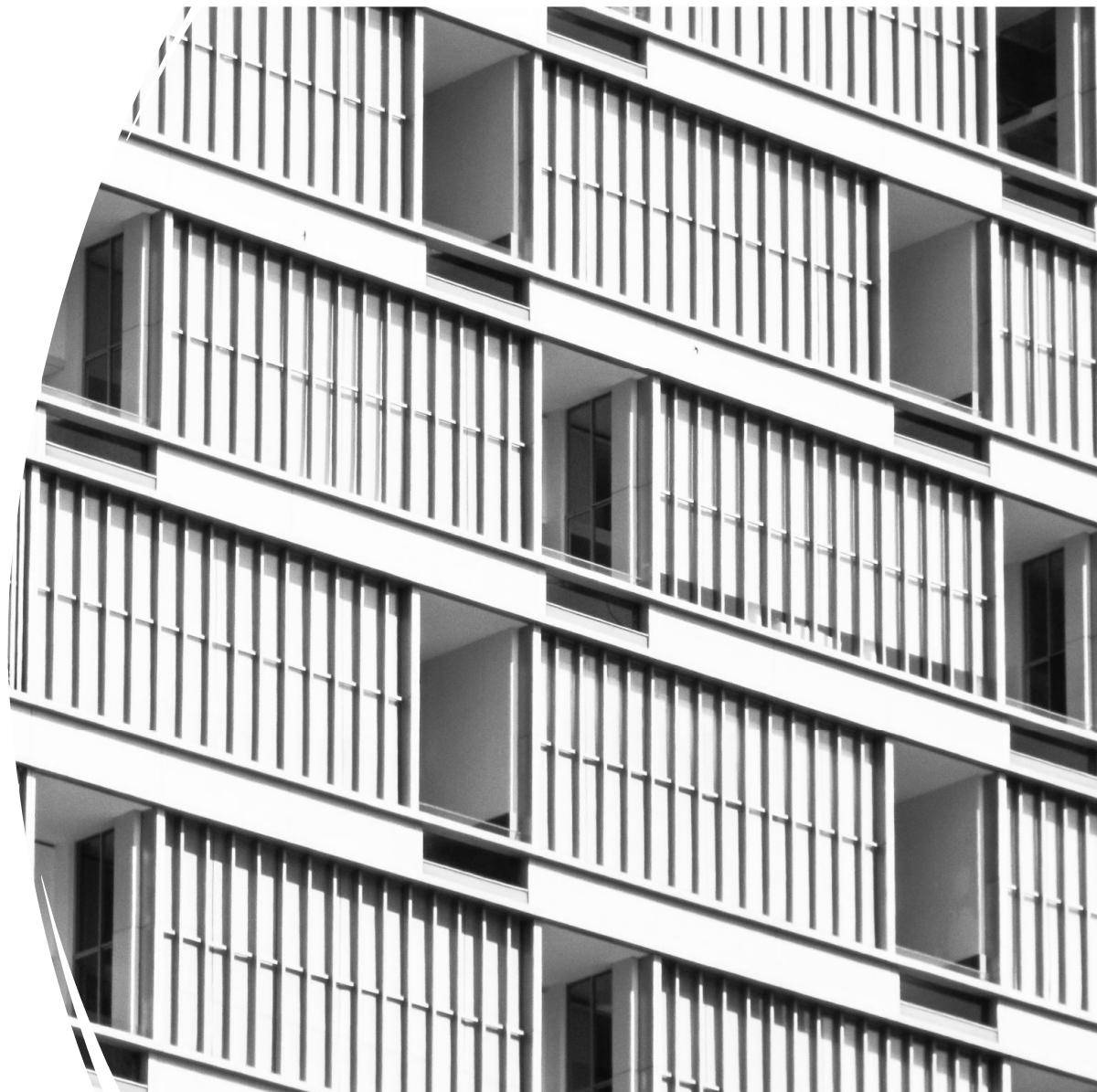
```
-bash-4.2$  
-hash-4.2$ hadoop version  
Hadoop 3.1.2  
Source code repository ssh://git@bitbucket.oci.oraclecorp.com:7999/bdcs/apache_bigtop.git -r 2ce21  
64edd68ab34a98a99cea5a1c2037a3a7ca1  
Compiled by root on 2024-07-30T16:09Z  
Compiled with protoc 2.5.0  
From source with checksum eaec305dfb8e514a567cd7fdbba59f0  
This command was run using /usr/odh/1.1.13/hadoop/hadoop-common-3.1.2.jar  
-bash-4.2$
```

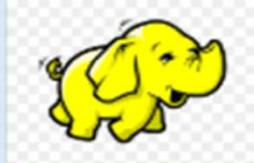
```
-bash-4.2$  
-bash-4.2$ yarn node -list all  
25/11/28 04:15:31 INFO client.RMProxy: Connecting to ResourceManager at bigdaimn0.sub03291929060.t  
rainingvcn.oraclevcn.com/10.1.0.251:8050  
25/11/28 04:15:32 INFO client.AHSProxy: Connecting to Application History server at bigdaiun0.sub0  
3291929060.trainingvcn.oraclevcn.com/10.1.0.73:10200  
Total Nodes:3  
      Node-Id          Node-State Node-Http-Address           Number-of-Running-Containers  
bigdaiwn2.sub03291929060.trainingvcn.oraclevcn.com:45454      RUNNING bigdaiwn2.sub03291  
929060.trainingvcn.oraclevcn.com:8042                      0          RUNNING bigdaiwn0.sub03291  
bigdaiwn0.sub03291929060.trainingvcn.oraclevcn.com:45454      0          RUNNING bigdaiwn1.sub03291  
929060.trainingvcn.oraclevcn.com:8042                      1          RUNNING bigdaiwn1.sub03291  
-bash-4.2$
```

- **Cluster Version:** Hadoop 3.1.2
- **Cluster Number of Nodes:** 3

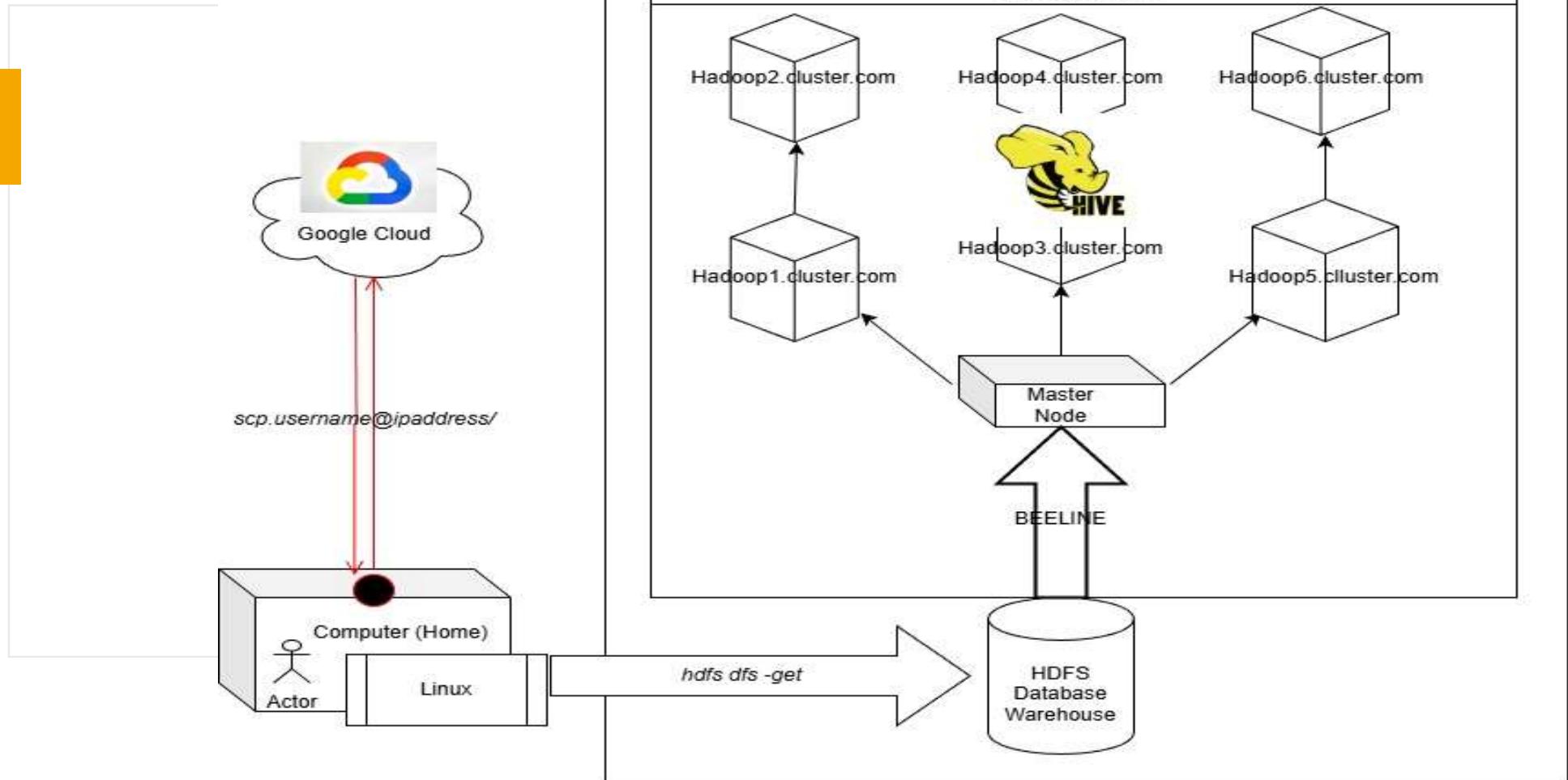


Architecture

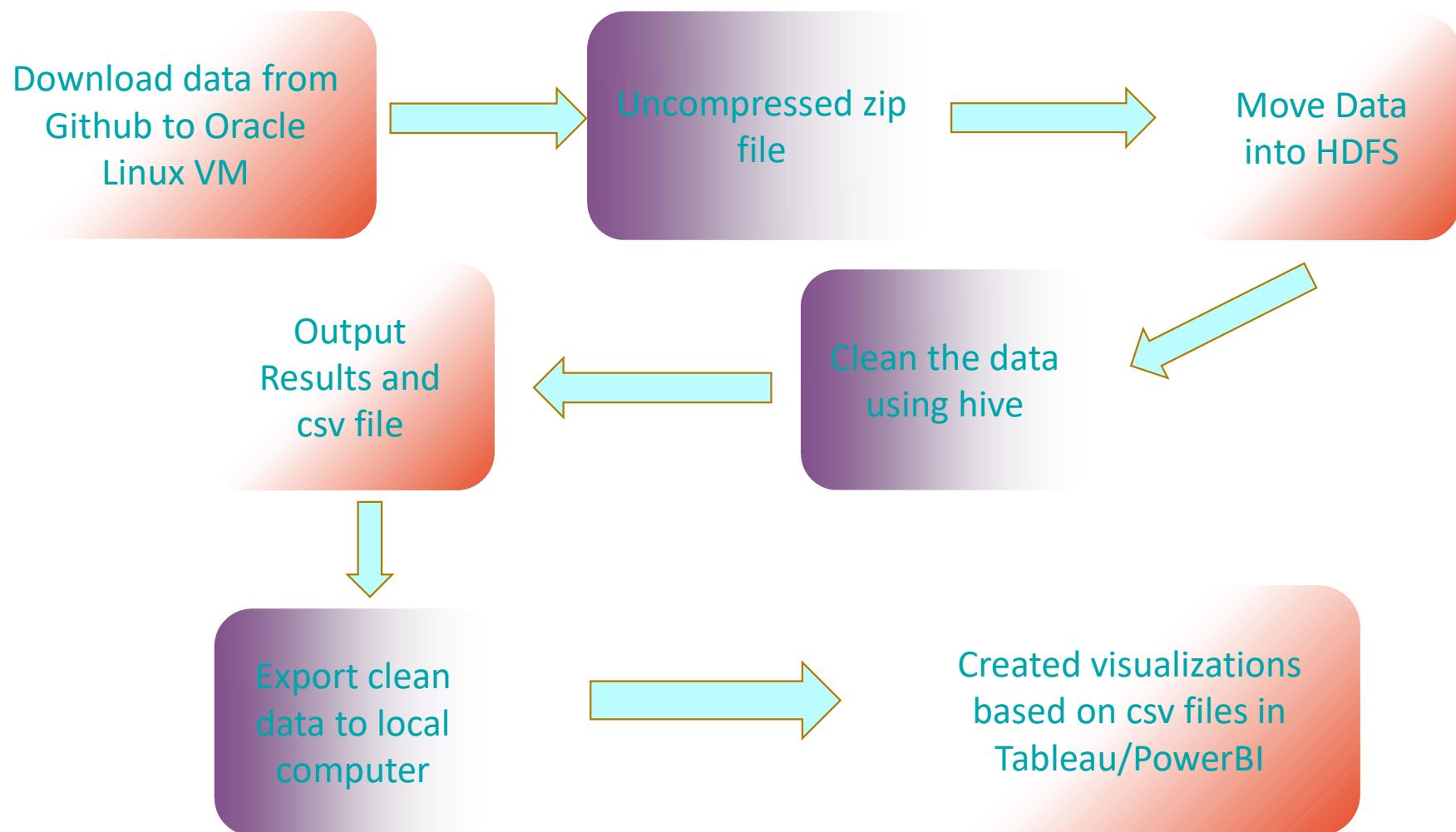


SOURCE	STAGING	DATA WAREHOUSE	QUERY & ANALYSIS	VISUALIZATION	REPOSITORY
 GOOGLE CLOUD (Source Code / Raw Data)	 <b>LINUX</b> Open Source O.S. Kernel	 HDFS HADOOP File System	 <b>APACHE HIVE</b>	 POWER BI	 <b>GITHUB</b>

# Hadoop Cluster Flowchart



# Workflow for Data Analysis



## Retrieving the Data

- The complete dataset was pulled from Github
- Wget command is used to download the data to Oracle server
- wget [https://github.com/chica-94/COVID19-Data-Analysis/raw/refs/heads/main/covid19\\_tables.zip](https://github.com/chica-94/COVID19-Data-Analysis/raw/refs/heads/main/covid19_tables.zip)

```
-bash-4.2$ wget -O covid19_aggregated.csv.gz https://storage.googleapis.com/covid19-open-data/v3/aggregated.csv.gz
--2025-11-28 03:46:01--  https://storage.googleapis.com/covid19-open-data/v3/aggregated.csv.gz
Resolving storage.googleapis.com (storage.googleapis.com)... 172.217.75.207, 192.178.164.207, 142.251.181.207, ...
Connecting to storage.googleapis.com (storage.googleapis.com)|172.217.75.207|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1420044994 (1.3G) [text/csv]
Saving to: 'covid19_aggregated.csv.gz'

100%[=====] 1,420,044,994 59.0MB/s   in 16s
2025-11-28 03:46:17 (85.7 MB/s) - 'covid19_aggregated.csv.gz' saved [1420044994/1420044994]

-bash-4.2$ ls -la covid19_aggregated.csv.gz
-rw-rw-r-- 1 bchica bchica 1420044994 Sep 16 2022 covid19_aggregated.csv.gz
-bash-4.2$ ls -lha covid19_aggregated.csv.gz
-rw-rw-r-- 1 bchica bchica 1.4G Sep 16 2022 covid19_aggregated.csv.gz
-bash-4.2$ |
```

1,420,044,994 59.0MB/s in 16s

Size of compressed folder (bytes)

# Extracting Dataset

- *Unzip the compressed folder*
- List all tables inside extracted folder

```
-bash-4.2$ unzip covid19_tables.zip
Archive: covid19_tables.zip
  creating: covid19_tables/
  inflating: covid19_tables/age.csv
  inflating: covid19_tables/demographics.csv
  inflating: covid19_tables/epidemiology.csv
  inflating: covid19_tables/gender.csv
  inflating: covid19_tables/geography.csv
  inflating: covid19_tables/health.csv
  inflating: covid19_tables/hospitalizations.csv
  inflating: covid19_tables/index.csv
  inflating: covid19_tables/vaccinations.csv
-bash-4.2$ ls -lha covid19_tables
total 942M
drwxrwxrwx  2 bchica bchica 4.0K Dec  3 04:41 .
drwx----- 10 bchica bchica 4.0K Dec  3 04:50 ..
-rw-rw-rw-  1 bchica bchica 212M Dec  3 00:14 age.csv
-rw-rw-rw-  1 bchica bchica 1.4M Dec  2 18:52 demographics.csv
-rw-rw-rw-  1 bchica bchica 457M Dec  2 20:17 epidemiology.csv
-rw-rw-rw-  1 bchica bchica 142M Dec  2 19:53 gender.csv
-rw-rw-rw-  1 bchica bchica 662K Dec  2 19:05 geography.csv
-rw-rw-rw-  1 bchica bchica  91K Dec  2 18:45 health.csv
-rw-rw-rw-  1 bchica bchica  55M Dec  2 18:41 hospitalizations.csv
-rw-rw-rw-  1 bchica bchica 1.1M Dec  2 18:42 index.csv
-rw-rw-rw-  1 bchica bchica  75M Dec  2 19:00 vaccinations.csv
-bash-4.2$
```

# Moving the Data

- Created a new directory in HDFS from the Linux environment to store the data.
- Uploaded the dataset into HDFS using the **-put** command.

```
----  
-bash-4.2$ cd covid19_tables/  
-bash-4.2$ hdfs dfs -mkdir /tmp/group3_covid19/  
-bash-4.2$ hdfs dfs -mkdir /tmp/group3_covid19/index  
-bash-4.2$ hdfs dfs -mkdir /tmp/group3_covid19/hospitalizations  
-bash-4.2$ hdfs dfs -mkdir /tmp/group3_covid19/health  
-bash-4.2$ hdfs dfs -mkdir /tmp/group3_covid19/vaccinations  
-bash-4.2$ hdfs dfs -mkdir /tmp/group3_covid19/demographics  
-bash-4.2$ hdfs dfs -mkdir /tmp/group3_covid19/geography  
-bash-4.2$ hdfs dfs -mkdir /tmp/group3_covid19/gender  
-bash-4.2$ hdfs dfs -mkdir /tmp/group3_covid19/epidemiology  
-bash-4.2$ hdfs dfs -mkdir /tmp/group3_covid19/age  
-bash-4.2$ hdfs dfs -mkdir /tmp/group3_covid19/gender  
mkdir: '/tmp/group3_covid19/gender': File exists  
-bash-4.2$ hdfs dfs -ls /tmp/group3_covid19/  
Found 9 items  
drwxr-xr-x - bchica hdfs 0 2025-12-03 04:56 /tmp/group3_covid19/age  
drwxr-xr-x - bchica hdfs 0 2025-12-03 04:56 /tmp/group3_covid19/demographics  
drwxr-xr-x - bchica hdfs 0 2025-12-03 04:56 /tmp/group3_covid19/epidemiology  
drwxr-xr-x - bchica hdfs 0 2025-12-03 04:56 /tmp/group3_covid19/gender  
drwxr-xr-x - bchica hdfs 0 2025-12-03 04:56 /tmp/group3_covid19/geography  
drwxr-xr-x - bchica hdfs 0 2025-12-03 04:56 /tmp/group3_covid19/health  
drwxr-xr-x - bchica hdfs 0 2025-12-03 04:56 /tmp/group3_covid19/hospitalizations  
drwxr-xr-x - bchica hdfs 0 2025-12-03 04:56 /tmp/group3_covid19/index  
drwxr-xr-x - bchica hdfs 0 2025-12-03 04:56 /tmp/group3_covid19/vaccinations  
-bash-4.2$ |
```

```
.bash-4.2$ hdfs dfs -put index.csv /tmp/group3_covid19/index  
-bash-4.2$ hdfs dfs -put hospitalizations.csv /tmp/group3_covid19/hospitalizations  
-bash-4.2$ hdfs dfs -put health.csv /tmp/group3_covid19/health  
-bash-4.2$ hdfs dfs -put vaccinations.csv /tmp/group3_covid19/vaccinations  
-bash-4.2$ hdfs dfs -put demographics.csv /tmp/group3_covid19/demographics  
-bash-4.2$ hdfs dfs -put geography.csv /tmp/group3_covid19/geography  
-bash-4.2$ hdfs dfs -put gender.csv /tmp/group3_covid19/gender  
-bash-4.2$ hdfs dfs -put age.csv /tmp/group3_covid19/age  
-bash-4.2$ hdfs dfs -put epidemiology.csv /tmp/group3_covid19/epidemiology  
-bash-4.2$ hdfs dfs -ls -R /tmp/group3_covid19/  
ls: '/tmp/group3_covid19/': No such file or directory  
-bash-4.2$ hdfs dfs -ls -R /tmp/group3_covid19/  
lrwxr-xr-x - bchica hdfs 0 2025-12-03 04:56 /tmp/group3_covid19/age  
rw-r--r-- 3 bchica hdfs 221791739 2025-12-03 04:59 /tmp/group3_covid19/age/age.csv  
lrwxr-xr-x - bchica hdfs 0 2025-12-03 04:59 /tmp/group3_covid19/demographics  
rw-r--r-- 3 bchica hdfs 1375489 2025-12-03 04:59 /tmp/group3_covid19/demographics/demographics.csv  
lrwxr-xr-x - bchica hdfs 0 2025-12-03 04:59 /tmp/group3_covid19/epidemiology  
rw-r--r-- 3 bchica hdfs 478654081 2025-12-03 04:59 /tmp/group3_covid19/epidemiology/epidemiology.csv  
lrwxr-xr-x - bchica hdfs 0 2025-12-03 04:59 /tmp/group3_covid19/gender  
rw-r--r-- 3 bchica hdfs 148284695 2025-12-03 04:59 /tmp/group3_covid19/gender/gender.csv  
lrwxr-xr-x - bchica hdfs 0 2025-12-03 04:59 /tmp/group3_covid19/geography  
rw-r--r-- 3 bchica hdfs 676897 2025-12-03 04:59 /tmp/group3_covid19/geography/geography.csv  
lrwxr-xr-x - bchica hdfs 0 2025-12-03 04:59 /tmp/group3_covid19/health  
rw-r--r-- 3 bchica hdfs 92483 2025-12-03 04:59 /tmp/group3_covid19/health/health.csv  
lrwxr-xr-x - bchica hdfs 0 2025-12-03 04:58 /tmp/group3_covid19/hospitalizations  
rw-r--r-- 3 bchica hdfs 56630483 2025-12-03 04:58 /tmp/group3_covid19/hospitalizations/hospitalizatio  
nrw-r--r-- 3 bchica hdfs 0 2025-12-03 04:58 /tmp/group3_covid19/index  
rw-r--r-- 3 bchica hdfs 1114818 2025-12-03 04:58 /tmp/group3_covid19/index/index.csv  
lrwxr-xr-x - bchica hdfs 0 2025-12-03 04:59 /tmp/group3_covid19/vaccinations  
rw-r--r-- 3 bchica hdfs 78235179 2025-12-03 04:59 /tmp/group3_covid19/vaccinations/vaccinations.csv  
.bash-4.2$ |
```

# Hive Data Processing (Preparing the data)

- Used Hive to clean, filter, and analyze large datasets
- Key Tasks:
  - Removed null/invalid rows
  - Aggregated data by ( country, month, years,...)
  - Calculated case fatality rates
  - Generated reduced tables for visualization
- Cleaned the data using OpenRefine and Hive
  - Final process dataset cleaned and reduced from 20.9 GB (Origin) -> 3.2 GB (MapReduce) -> 2.1 MB (Visualization)

# Obtain datasets via Google COVID-19 Open Data

## Download the csv files of the following tables:

- Index
- Demographics
- Epidemiology
- Geography
- Health
- Hospitalizations
- Vaccinations
- By age
- By sex

Total size: ~21 GB

The screenshot shows the Google COVID-19 Open Data Repository. At the top, there's a navigation bar with links for 'COVID-19 Open Data', 'Data visualizer', and 'raw data'. Below the navigation, a message says 'As of September 15, 2022, we have turned off real-time updates in this repository.' A 'Learn more' link is provided. The main content area is titled 'COVID-19 Open Data Repository'. It features a central graphic of a globe surrounded by various icons representing data types like graphs, charts, and documents. Below the graphic, a section titled 'Dive into the data' explains the purpose of the repository. Two main options are presented: 'Visualize the data' (with a link to a visualization tool) and 'Access the raw dataset' (with a link to download the dataset in CSV or JSON format). A note at the bottom encourages users to contact support if they can't find what they're looking for.

Table	Indexed by*	Content	Source*	Download
<a href="#">Ancestry</a>	[key] [date]	Flat, compressed table with records from (almost) all other tables joined by date and/or key; see below for more details	All tables below	<a href="#">CSV</a>
<a href="#">Index</a>	[key]	Various names and codes, useful for joining with other datasets	Wikidata, DataCommons, Eurostat	<a href="#">CSV</a> <a href="#">JSON</a>
<a href="#">Demographics</a>	[key]	Various (current*) population statistics	Wikidata, DataCommons, WorldBank, WorldPop, Eurostat	<a href="#">CSV</a> <a href="#">JSON</a>
<a href="#">Economy</a>	[key]	Various (current*) economic indicators	Wikidata, DataCommons, Eurostat	<a href="#">CSV</a> <a href="#">JSON</a>
<a href="#">Epidemiology</a>	[key] [date]	COVID-19 cases, deaths, recoveries, and tests	Various*	<a href="#">CSV</a> <a href="#">JSON</a>
<a href="#">Emergency declarations</a>	[key] [date]	Government emergency declarations and mitigation policies	LawAtlas Project	<a href="#">CSV</a>
<a href="#">Geography</a>	[key]	Geographical information about the region	Wikidata	<a href="#">CSV</a> <a href="#">JSON</a>
<a href="#">Health</a>	[key]	Health indicators for the region	Wikidata, WorldBank, Eurostat	<a href="#">CSV</a> <a href="#">JSON</a>
<a href="#">Hospitalizations</a>	[key] [date]	Information related to patients of COVID-19 and hospitals	Various*	<a href="#">CSV</a> <a href="#">JSON</a>
<a href="#">Mobility</a>	[key] [date]	Various metrics related to the movement of people. To download or use the data, you must agree to the <a href="#">Google Terms of Service</a> .	Google	<a href="#">CSV</a> <a href="#">JSON</a>
<a href="#">Search trends</a>	[key] [date]	Trends in symptom search volumes due to COVID-19. To download or use the data, you must agree to the <a href="#">Google Terms of Service</a> .	Google	<a href="#">CSV</a>
<a href="#">Vaccination access</a>	[place_id]	Metrics quantifying access to COVID-19 vaccination sites. To download or use the data, you must agree to the <a href="#">Google Terms of Service</a> .	Google	<a href="#">CSV</a>
<a href="#">Vaccination search</a>	[key] [date]	Trends in Google searches for COVID-19 vaccination information. To download or use the data, you must agree to the <a href="#">Google Terms of Service</a> .	Google	<a href="#">CSV</a>
<a href="#">Vaccinations</a>	[key] [date]	Trends in persons vaccinated and popular vaccination rate. Requires various Covid-19 vaccines. To download or use the data, you must agree to the <a href="#">Google Terms of Service</a> .	Google	<a href="#">CSV</a>
<a href="#">Government response</a>	[key] [date]	Government interventions and their relative stringency	University of Oxford	<a href="#">CSV</a> <a href="#">JSON</a>
<a href="#">Weather</a>	[key] [date]	Dated meteorological information for each region	NOAA	<a href="#">CSV</a>
<a href="#">WorldBank</a>	[key]	Latest record for each indicator from WorldBank for all reporting countries	WorldBank	<a href="#">CSV</a> <a href="#">JSON</a>
<a href="#">By age</a>	[key] [date]	Epidemiology and hospitalizations data stratified by age	Various*	<a href="#">CSV</a> <a href="#">JSON</a>
<a href="#">By sex</a>	[key] [date]	Epidemiology and hospitalizations data stratified by sex	Various*	<a href="#">CSV</a> <a href="#">JSON</a>

# Cleaning up: (OpenRefine)

Re-order / Remove columns

Drag columns to re-order      Drop columns here to remove

location_key	openstreetmap_id
latitude	elevation_m
longitude	area_sq_km
	area_rural_sq_km
	area_urban_sq_km

Remove all   Add all      OK   Cancel

12/14/2025

OpenRefine geography csv [Permalink](#)

> 22,130 rows

Show as: [rows](#) [records](#)   Show: [5](#) [10](#) [25](#) [50](#) [100](#) [500](#) [1000](#) rows

All	location_key	openstreetmap_id	latitude	longitude	elevation_m	area_sq_km	area_rural_sq_km	area_urban_sq_km
1.	AD	9407	42.558333	1.555278		470		
2.	AE	307763	24.4	54.3		83600	70575	8568
3.	AF	303427	33.0	66.0		652860		
4.	AF_BAL	1674795	36.7	67.116667	340	16186		
5.	AF_BAM		34.75	67.25	3042	14175		
6.	AF_BDG		35.0	63.75	1589	20591		
7.	AF_BDS		38.0	71.0	3669	44059		
8.	AF_BGL		35.87	68.93	2013	21118		
9.	AF_DAY		33.75	66.25	2200	18088		
10.	AF_FRA	1674802	32.5	63.5	1081	48470		

OpenRefine geography csv [Permalink](#)

> 22,130 rows

Show as: [rows](#) [records](#)   Show: [5](#) [10](#) [25](#) [50](#) [100](#) [500](#) [1000](#) rows

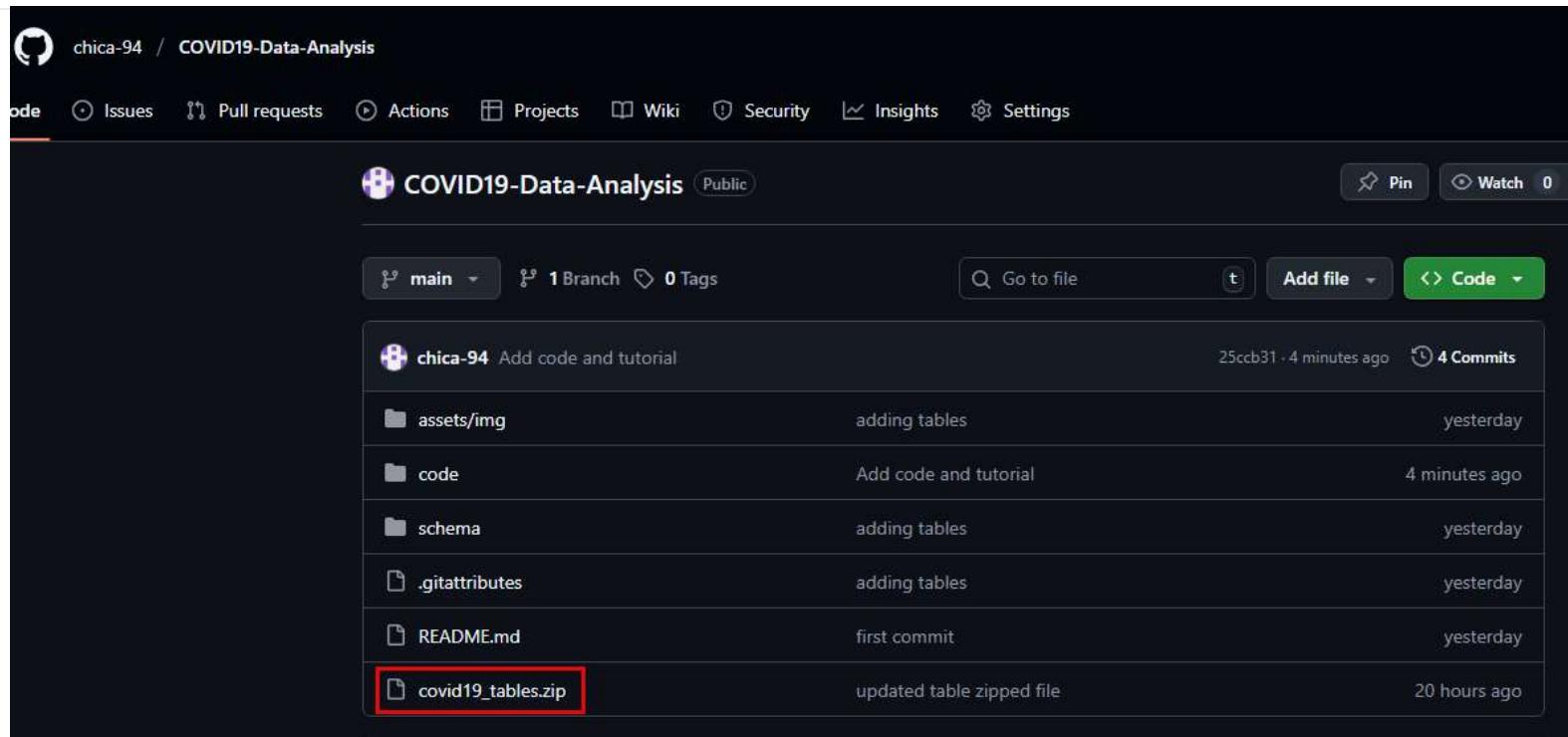
All	location_key	latitude	longitude
1.	AD	42.558333	1.555278
2.	AE	24.4	54.3
3.	AF	33.0	66.0
4.	AF_BAL	36.7	67.116667
5.	AF_BAM	34.75	67.25
6.	AF_BDG	35.0	63.75
7.	AF_BDS	38.0	71.0
8.	AF_BGL	35.87	68.93
9.	AF_DAY	33.75	66.25
10.	AF_FRA	32.5	63.5

OpenRefine project archive to file  
base ▾  
last »

Export ▾

- Tab-separated value
- Comma-separated value** (highlighted)
- HTML table
- Excel (.xls)
- Excel 2007+ (.xlsx)
- ODF spreadsheet
- Custom tabular...
- SQL...
- Templating...
- Wikibase edits...
- QuickStatements file
- Wikibase schema

# Added cleaned tables to Github repo



The screenshot shows a GitHub repository page for 'chica-94 / COVID19-Data-Analysis'. The repository is public. The commit history is displayed, showing the following commits:

Commit	Message	Time
chica-94 Add code and tutorial	adding tables	yesterday
code	Add code and tutorial	4 minutes ago
schema	adding tables	yesterday
.gitattributes	adding tables	yesterday
README.md	first commit	yesterday
covid19_tables.zip	updated table zipped file	20 hours ago

# Creating Tables: Apache Hive

```
0: jdbc:hive2://bigdatuno.sub03291929060.trai> DROP TABLE IF EXISTS covid19_index;
INFO : Compiling command(queryId=hive_20251203051348_e7af546c-0846-4725-9270-9cb0f3250858); DRO
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldschemas:null, properties:null)
INFO : Completed compiling command(queryId=hive_20251203051348_e7af546c-0846-4725-9270-9cb0f3250858); DRO
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20251203051348_e7af546c-0846-4725-9270-9cb0f3250858); DRO
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20251203051348_e7af546c-0846-4725-9270-9cb0f3250858); DRO
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.363 seconds)
0: jdbc:hive2://bigdatuno.sub03291929060.trai> CREATE EXTERNAL TABLE IF NOT EXISTS covid19_index
    (. . . . . > location_key STRING, country_code STRING,country_
name STRING)
    (. . . . . > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    . . . . . > STORED AS TEXTFILE LOCATION '/tmp/group3_covid19/'
    . . . . . > TBLPROPERTIES ("skip.header.line.count"="1");
INFO : Compiling command(queryId=hive_20251203051356_da784a9d-7997-4838-aa10-1c503e205ca0); CRE
(
location_key STRING, country_code STRING,country_name STRING,subregion1_name STRING,subregion2_name STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION '/tmp/group3_covid19/index/'
TBLPROPERTIES ("skip.header.line.count"="1")
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldschemas:null, properties:null)
INFO : Completed compiling command(queryId=hive_20251203051356_da784a9d-7997-4838-aa10-1c503e205ca0); Time taken: 0.014 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20251203051356_da784a9d-7997-4838-aa10-1c503e205ca0); CREATE EXTERNAL TABLE IF NOT EXISTS covid19_index
(
location_key STRING, country_code STRING,country_name STRING,subregion1_name STRING,subregion2_name STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION '/tmp/group3_covid19/index/'
TBLPROPERTIES ("skip.header.line.count"="1")
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20251203051356_da784a9d-7997-4838-aa10-1c503e205ca0); Time taken: 0.127 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.146 seconds)
```

# HiveQL: Select Query (index)

```
0 rows affected (0.146 seconds)
0: jdbc:hive2://bigdaiun0.sub03291929060.trai> SELECT * FROM covid19_index LIMIT 10;
INFO : Compiling command(queryId=hive_20251203051403_41ed71c1-efa3-4a50-9f41-fcb9c6ab137b): SELECT * FROM covid19_index LIMIT 10
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retryal = false)
INFO : Returning Hive schema: Schema(fieldschemas:[FieldSchema(name:covid19_index.location_key, type:string, comment:null), FieldSchema(name:covid19_index.country_code, type:string, comment:null), FieldSchema(name:covid19_index.country_name, type:string, comment:null), FieldSchema(name:covid19_index.subregion1_name, type:string, comment:null), FieldSchema(name:covid19_index.subregion2_name, type:string, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20251203051403_41ed71c1-efa3-4a50-9f41-fcb9c6ab137b); Time taken: 0.218 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20251203051403_41ed71c1-efa3-4a50-9f41-fcb9c6ab137b): SELECT * FROM covid19_index LIMIT 10
INFO : Completed executing command(queryId=hive_20251203051403_41ed71c1-efa3-4a50-9f41-fcb9c6ab137b); Time taken: 0.0 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+-----+-----+-----+
| covid19_index.location_key | covid19_index.country_code | covid19_index.country_name | covid19_index.subregion1_name | covid19_index.subregion2_name |
+-----+-----+-----+-----+
| AD | AD | Andorra | | |
| AE | AE | United Arab Emirates | | |
| AF | AF | Afghanistan | | |
| AF_BAL | AF | Afghanistan | Balkh | |
| AF_BAM | AF | Afghanistan | Bamyan | |
| AF_BDG | AF | Afghanistan | Badghis | |
| AF_BDS | AF | Afghanistan | Badakhshan | |
| AF_BGL | AF | Afghanistan | Baghlan | |
| AF_DAY | AF | Afghanistan | Daykundi | |
| AF_FRA | AF | Afghanistan | Farah | |
+-----+-----+-----+-----+
0 rows selected (0.232 seconds)
```

12/14/2025

20

# HiveQL: Select Query (health)

```
0: jdbc:hive2://bigdaiuno.sub03291929060.trai> SELECT * FROM covid19_health LIMIT 10;
INFO : Compiling command(queryId=hive_20251203052123_6b9cd0dd-a2f7-435d-9a12-0426a98ee287): SELECT * FROM covid19_health LIMIT 10
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldschemas:[FieldSchema(name:covid19_health.location_key, type:string, comment:null), FieldSchema(name:covid19_health.adult_male_mortality_rate, type:double, comment:null), FieldSchema(name:covid19_health.adult_female_mortality_rate, type:double, comment:null), FieldSchema(name:covid19_health.life_expectancy, type:double, comment:null), FieldSchema(name:covid19_health.diabetes_prevalence, type:double, comment:null), FieldSchema(name:covid19_health.health_expenditure_usd, type:double, comment:null), FieldSchema(name:covid19_health.out_of_pocket_health_expenditure_usd, type:double, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20251203052123_6b9cd0dd-a2f7-435d-9a12-0426a98ee287); Time taken: 0.22 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20251203052123_6b9cd0dd-a2f7-435d-9a12-0426a98ee287): SELECT * FROM covid19_health LIMIT 10
INFO : Completed executing command(queryId=hive_20251203052123_6b9cd0dd-a2f7-435d-9a12-0426a98ee287); Time taken: 0.0 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+-----+-----+-----+-----+
| covid19_health.location_key | covid19_health.adult_male_mortality_rate | covid19_health.adult_female_mortality_rate | covid19_health.life_expectancy | covid19_health.diabetes_prevalence | covid19_health.health_expenditure_usd | covid19_health.out_of_pocket_health_expenditure_usd |
+-----+-----+-----+-----+-----+
| AD | NULL | NULL | NULL | 7.7
| AD | 4040.786621 | 1688.12146 | 44.863 | 77.814 | 16.3
| AE | 69.555 | 256.034485 | 192.532 | 64.486 | 9.2
| AF | 1357.017456 | 237.554 | 50.665913 | 83.136 | 13.1
| AF | 237.554 | 50.665913 | 83.136 | 76.885 | 13.1
| AG | 67.12265 | 673.85968 | 235.749039 | 49.486 | 9.0
| AL | 126.917 | 93.315 | NULL | 78.9 | 9.0
| AL | 93.315 | NULL | 78.9 | 74.945 | 6.1
| AM | 673.85968 | 173.428 | 65.595 | 220.291 | 21
| AM | 173.428 | 407.635864 | 343.832977 | 220.291 | 60.782 | 4.5
| AO | 93.315 | 327.044 | 220.291 | 60.782 | 4.5
```

# HiveQL: Select Query (geography)

```
0: jdbc:hive2://bigdatuno.sub03291929060.traj>
0: jdbc:hive2://bigdatuno.sub03291929060.traj>   SELECT * FROM covid19_geography LIMIT 10;
INFO  : Compiling command(queryId=hive_20251203052620_5b9b836e-9d35-46cf-aed2-97447e3613c2): SELECT * FROM covid19_geography LIMIT 10
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldschemas:[FieldSchema(name:covid19_geography.location_key, type:string, comment:null), FieldSchema(name:covid19_geography.latitude, type:double, comment:null), FieldSchema(name:covid19_geography.longitude, type:double, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20251203052620_5b9b836e-9d35-46cf-aed2-97447e3613c2); Time taken: 0.232 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20251203052620_5b9b836e-9d35-46cf-aed2-97447e3613c2): SELECT * FROM covid19_geography LIMIT 10
INFO  : Completed executing command(queryId=hive_20251203052620_5b9b836e-9d35-46cf-aed2-97447e3613c2); Time taken: 0.0 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-----+-----+-----+
| covid19_geography.location_key | covid19_geography.latitude | covid19_geography.longitude |
+-----+-----+-----+
| AD                         | 42.558333              | 1.555278
| AE                         | 24.4                   | 54.3
| AF                         | 33.0                   | 66.0
| AF_BAL                     | 36.7                   | 67.116667
| AF_BAM                     | 34.75                  | 67.25
| AF_BDG                     | 35.0                   | 63.75
| AF_BDS                     | 38.0                   | 71.0
| AF_BGL                     | 35.87                  | 68.93
| AF_DAY                     | 33.75                  | 66.25
| AF_FRA                     | 32.5                   | 63.5
+-----+-----+-----+
10 rows selected (0.25 seconds)
```

# HiveQL: Insert Overwrite (aggregated)

```
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20251203060935_5be0bfcd-67f4-4ee4-b156-90e4dc59ff00): INSERT OVERWRITE TABLE covid19_aggregated
SELECT epi.event_date,epi.location_key, idx.country_code, idx.country_name, idx.subregion1_name, idx.subregion2_name,epi.new_confirmed,epi.cumulative_deceased,epi.cumulative_deceased,epi.new_recovered,epi.cumulative_recovered,hosp.new_hospitalized_patients,hosp.cumulative_hospitalized_patients,hosp.cumulative_intensive_care_patients,vac.new_persons_fully_vaccinated,vac.cumulative_persons_fully_vaccinated,health.adult_adult_female_mortality_rate,health.life_expectancy,health.diabetes_prevalence,health.health_expenditure_usd,health.out_of_pocket_healthcare_population,demo.population_male,demo.population_female,demo.population_age_00_09,demo.population_age_10_19,demo.population_age_20_29,demo.population_age_40_49,demo.population_age_50_59,demo.population_age_60_69,demo.population_age_70_79,demo.population_age_80_and_older,geo.latitude,geo.longitude,geo.deceased_male,gender.cumulative_deceased_male,gender.new_deceased_female,gender.cumulative_deceased_female,gender.new_recovered_male,gender.new_recovered_female,gender.cumulative_recovered_female,age.new_recovered_age_0_9,age.new_recovered_age_10_19,age.new_recovered_age_20_29,age.new_recovered_age_30_39,age.new_recovered_age_40_49,age.new_recovered_age_50_59,age.new_recovered_age_60_69,age.new_recovered_age_70_79,age.new_recovered_age_90_99,age.cumulative_recovered_age_90_99
FROM covid19_epidemiology epi
LEFT JOIN covid19_index idx ON epi.location_key = idx.location_key
LEFT JOIN covid19_hospitalizations hosp ON epi.location_key = hosp.location_key AND epi.event_date = hosp.event_date
LEFT JOIN covid19_vaccinations vac ON epi.location_key = vac.location_key AND epi.event_date = vac.event_date
LEFT JOIN covid19_health health ON epi.location_key = health.location_key
LEFT JOIN covid19_demographics demo ON epi.location_key = demo.location_key
LEFT JOIN covid19_geography geo ON epi.location_key = geo.location_key
LEFT JOIN covid19_gender gender ON epi.location_key = gender.location_key AND epi.event_date = gender.event_date
LEFT JOIN covid19_age age ON epi.location_key = age.location_key AND epi.event_date = age.event_date
INFO : Query ID = hive_20251203060935_5be0bfcd-67f4-4ee4-b156-90e4dc59ff00
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Subscribed to counters: [] for queryId: hive_20251203060935_5be0bfcd-67f4-4ee4-b156-90e4dc59ff00
INFO : Tez session hasn't been created yet. Opening session
INFO : Dag name: INSERT OVERWRITE TABLE covid19_aggregated (Stage-1)
INFO : Setting tez.task.scale.memory.reserve-fraction to 0.30000001192092896
INFO : Status: Running (Executing on YARN cluster with App id application_1756243379751_1599)

-----  

      VERTICES    MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

-----  

Map 5 ..... container SUCCEEDED 1 1 0 0 0 0  

Map 1 ..... container RUNNING 1 0 1 0 0 0  

Map 6 ..... container RUNNING 1 0 1 0 0 0  

Map 7 ..... container RUNNING 1 0 1 0 0 0  

Reducer 2 ..... container INITED 130 0 0 130 0 0  

Map 8 ..... container SUCCEEDED 1 1 0 0 0 0  

Map 9 ..... container SUCCEEDED 1 1 0 0 0 0  

Map 10 ..... container SUCCEEDED 1 1 0 0 0 0  

Reducer 3 ..... container INITED 226 0 0 226 0 0  

Map 11 ..... container RUNNING 1 0 1 0 0 0  

Map 12 ..... container RUNNING 1 0 1 0 0 0  

Reducer 4 ..... container INITED 818 0 0 818 0 0  

-----  

VERTICES: 04/12 [>>-----] 0% ELAPSED TIME: 5.82 s  

-----
```

# Completed Tables

```
b: jdbc:hive2://bigdaiun0.sub03291929060.traj> show tables;  
INFO : Compiling command(queryId=hive_20251203061324_7d68ac0c-b  
INFO : Concurrency mode is disabled, not creating a lock manager  
INFO : Semantic Analysis Completed (retrial = false)  
INFO : Returning Hive schema: Schema(fieldschemas:[FieldSchema(name:col_name  
INFO : from deserializer), FieldSchema(name:comment, type:string, comment:from  
INFO : Completed compiling command(queryId=hive_20251203061459_bbe78cfe-2de0  
INFO : Concurrency mode is disabled, not creating a lock manager  
INFO : Executing command(queryId=hive_20251203061459_bbe78cfe-2de0-4b94-b687  
INFO : Starting task [Stage-0:DDL] in serial mode  
INFO : Completed executing command(queryId=hive_20251203061459_bbe78cfe-2de0  
INFO : OK  
INFO : Concurrency mode is disabled, not creating a lock manager  
+-----+-----+  
| tab_name |  
+-----+  
| covid19_age |  
| covid19_aggregated |  
| covid19_demographics |  
| covid19_epidemiology |  
| covid19_gender |  
| covid19_geography |  
| covid19_health |  
| covid19_hospitalizations |  
| covid19_index |  
| covid19_vaccinations |  
+-----+  
| ts |  
+-----+
```

12/14/2025

```
b: jdbc:hive2://bigdaiun0.sub03291929060.traj> DESCRIBE covid19_aggregated;  
INFO : Compiling command(queryId=hive_20251203061459_bbe78cfe-2de0-4b94-b687  
INFO : Concurrency mode is disabled, not creating a lock manager  
INFO : Semantic Analysis Completed (retrial = false)  
INFO : Returning Hive schema: Schema(fieldschemas:[FieldSchema(name:col_name  
INFO : from deserializer), FieldSchema(name:comment, type:string, comment:from  
INFO : Completed compiling command(queryId=hive_20251203061459_bbe78cfe-2de0  
INFO : Concurrency mode is disabled, not creating a lock manager  
INFO : Executing command(queryId=hive_20251203061459_bbe78cfe-2de0-4b94-b687  
INFO : Starting task [Stage-0:DDL] in serial mode  
INFO : Completed executing command(queryId=hive_20251203061459_bbe78cfe-2de0  
INFO : OK  
INFO : Concurrency mode is disabled, not creating a lock manager  
+-----+-----+-----+  
| col_name | data_type | comment |  
+-----+-----+-----+  
| event_date | date |  
| location_key | string |  
| country_code | string |  
| country_name | string |  
| subregion1_name | string |  
| subregion2_name | string |  
| new_confirmed | int |  
| cumulative_confirmed | int |  
| new_deceased | int |  
| cumulative_deceased | int |  
| new_recovered | int |  
| cumulative_recovered | int |  
| new_hospitalized_patients | int |  
| cumulative_hospitalized_patients | int |  
| new_intensive_care_patients | int |  
| cumulative_intensive_care_patients | int |  
| new_persons_fully_vaccinated | int |  
| cumulative_persons_fully_vaccinated | int |  
| adult_male_mortality_rate | double |  
| adult_female_mortality_rate | double |  
| life_expectancy | double |  
| diabetes_prevalence | double |  
| health_expenditure_usd | double |  
| out_of_pocket_health_expenditure_usd | double |  
| population | int |  
| population_male | int |  
| population_female | int |  
| population_age_00_09 | int |  
| population_age_10_19 | int |  
| population_age_20_29 | int |  
| population_age_30_39 | int |  
| population_age_40_49 | int |  
| population_age_50_59 | int |  
| population_age_60_69 | int |  
| population_age_70_79 | int |  
| population_age_80_and_older | int |  
| latitude | double |  
| longitude | double |  
| new_deceased_male | int |  
| cumulative_deceased_male | int |  
| new_deceased_female | int |  
| cumulative_deceased_female | int |  
| new_recovered_male | int |  
| cumulative_recovered_male | int |  
| new_recovered_female | int |  
| cumulative_recovered_female | int |  
| new_recovered_age_0_9 | int |  
| new_recovered_age_10_19 | int |  
| new_recovered_age_20_29 | int |  
| new_recovered_age_30_39 | int |  
| new_recovered_age_40_49 | int |
```

24

# Get the top 10 countries

```
SELECT country_name,MAX(cumulative_confirmed)
AS total_cases
FROM covid19_aggregated
WHERE cumulative_confirmed IS NOT NULL
GROUP BY country_name
ORDER BY total_cases DESC
LIMIT 10;
```

country_name	total_cases
United States of America	92440495
India	44516479
Brazil	34568833
France	33766090
Germany	32604993
South Korea	24264470
United Kingdom	23554971
Italy	22114423
Russia	20265004
Japan	19868288

# Create top 10 table

- Create an empty table called **covid19\_top10\_countries**, which will contain the cumulative confirmed cases for each country and its subregions (city & states) from **covid19\_aggregated** tables once the data has been formatted.
- It reduced the number of records from 12 million to 10 thousand
  - `SELECT COUNT(*) AS covid19_records FROM covid19_aggregated;`
  - `SELECT COUNT(*) AS covid19_top10_records FROM covid19_top10_countries;`
- The size reduced from 3.2GB to 2.1 MB

```
+-----+  
| covid19_records |  
+-----+  
| 12525825 |  
+-----+
```

```
+-----+  
| covid19_top10_records |  
+-----+  
| 10660 |  
+-----+
```

```
-bash-4.2$ du -h covid19_*
```

3.2G	covid19_aggregated.csv
2.1M	covid19_top10_countries.csv

```
-bash-4.2$ |
```

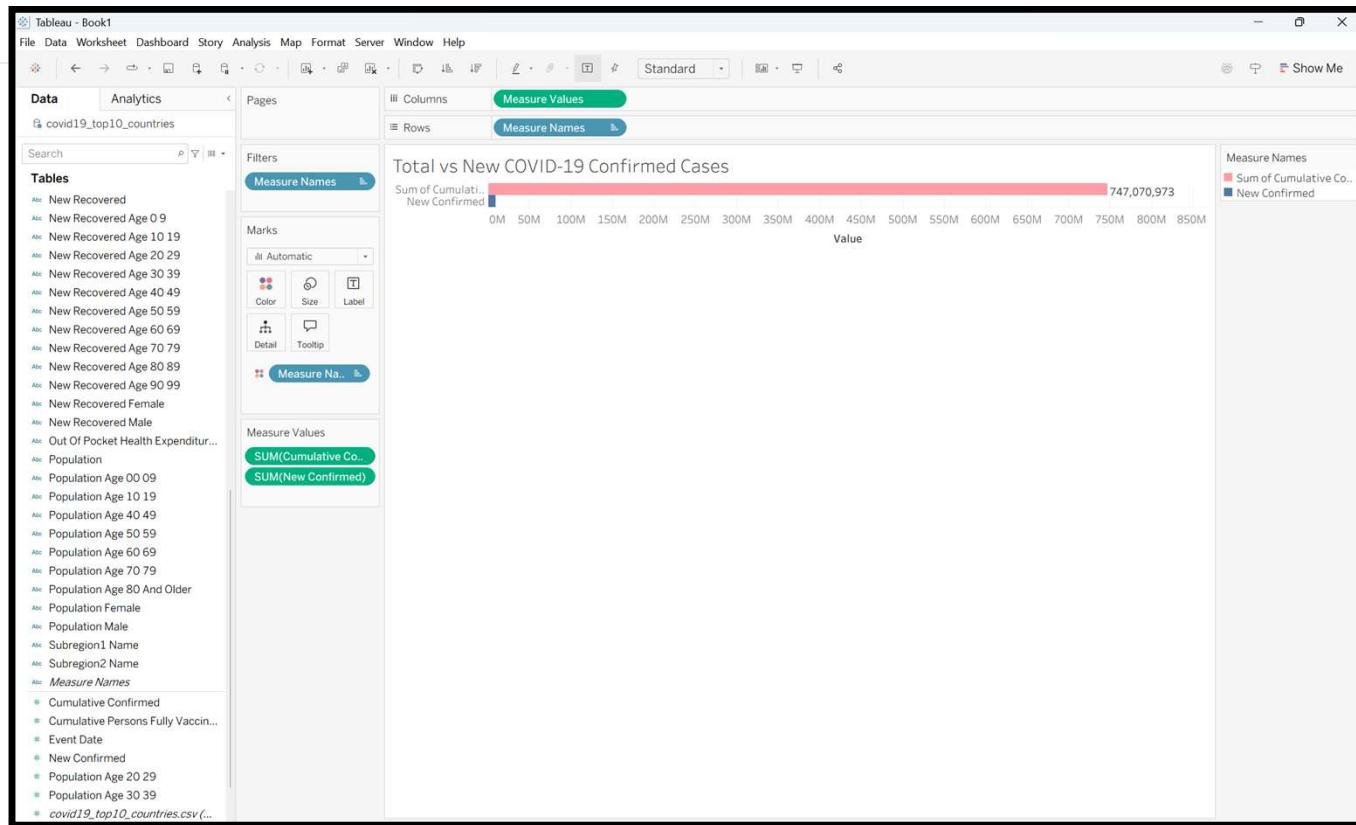
# CSV → Visualization

```
bash-4.2$ ls
ge.csv demographics.csv epidemiology.csv gender.csv geography.csv health.csv hospitalizations.csv index.csv vaccinations.csv
bash-4.2$ hdfs dfs -getmerge /tmp/group3_covid19/aggregated/000* covid19_aggregated.csv
bash-4.2$ ls covid19_aggregated.csv
covid19_aggregated.csv
bash-4.2$ ls -lha covid19_aggregated.csv
rw-r--r-- 1 bchica bchica 3.2G Dec  3 06:20 covid19_aggregated.csv
```

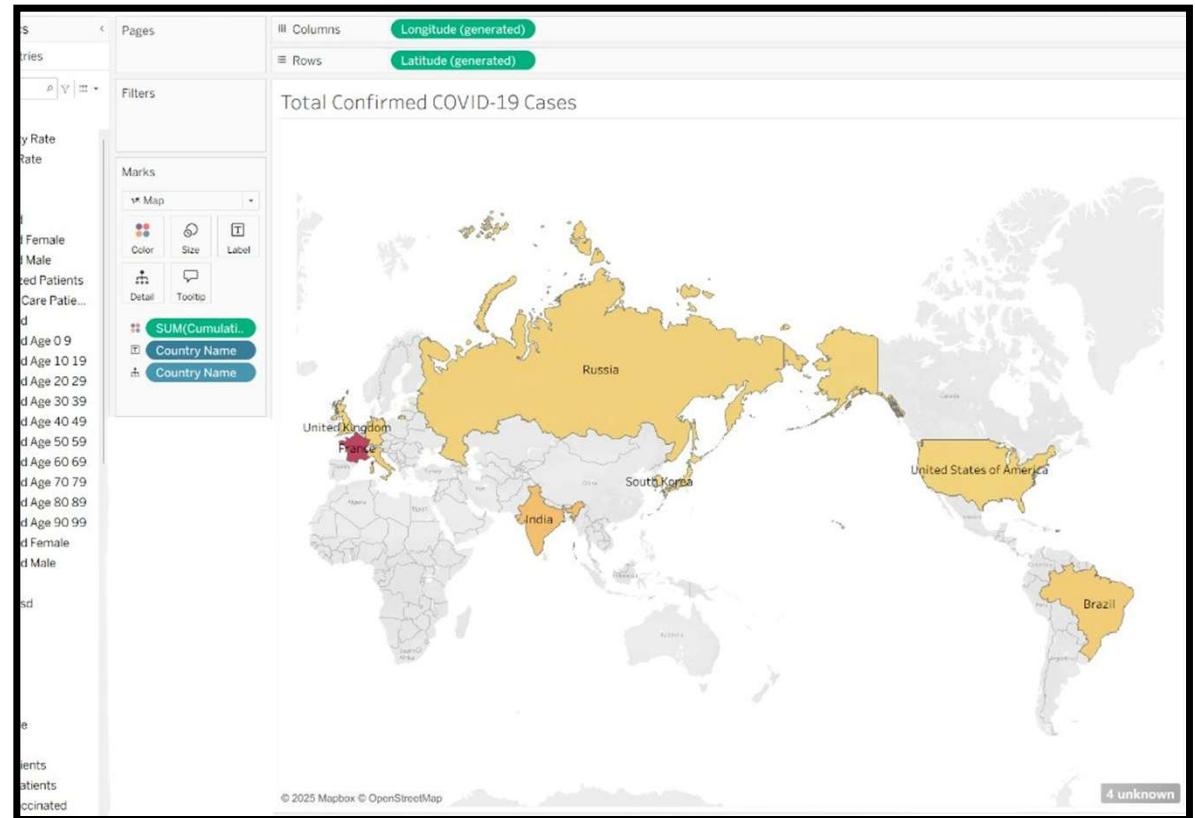
The screenshot shows a Microsoft Excel spreadsheet titled "COVID-19-geographic-disbtribution-worldwide-2020-12-14(COVID-19-geograph...)" with the file path "C:\Users\bchica\Downloads\covid19\_aggregated.csv". The table has 22 rows and 14 columns. The columns are labeled: dateRep, day, month, year, cases, deaths, countriesAndTerritories, geol, countryterritoryCode, popData2019, continentExp, and Cumulative\_number\_for\_14\_days\_of\_COVID. The data shows daily COVID-19 statistics for Afghanistan from December 1, 2019, to December 20, 2020. All values for cases, deaths, and the cumulative number are zero. The continentExp column consistently shows "Asia". The last row (row 22) is highlighted with a green border. The status bar at the bottom left shows "12/14/2025".

dateRep	day	month	year	cases	deaths	countriesAndTerritories	geol	countryterritoryCode	popData2019	continentExp	Cumulative_number_for_14_days_of_COVID
12/31/2019	31	12	2019	0	0	Afghanistan	AF	AFG		38041757	Asia
1/1/2020	1	1	2020	0	0	Afghanistan	AF	AFG		38041757	Asia
1/2/2020	2	1	2020	0	0	Afghanistan	AF	AFG		38041757	Asia
1/3/2020	3	1	2020	0	0	Afghanistan	AF	AFG		38041757	Asia
1/4/2020	4	1	2020	0	0	Afghanistan	AF	AFG		38041757	Asia
1/5/2020	5	1	2020	0	0	Afghanistan	AF	AFG		38041757	Asia
1/6/2020	6	1	2020	0	0	Afghanistan	AF	AFG		38041757	Asia
1/7/2020	7	1	2020	0	0	Afghanistan	AF	AFG		38041757	Asia
1/8/2020	8	1	2020	0	0	Afghanistan	AF	AFG		38041757	Asia
1/9/2020	9	1	2020	0	0	Afghanistan	AF	AFG		38041757	Asia
1/10/2020	10	1	2020	0	0	Afghanistan	AF	AFG		38041757	Asia
1/11/2020	11	1	2020	0	0	Afghanistan	AF	AFG		38041757	Asia
1/12/2020	12	1	2020	0	0	Afghanistan	AF	AFG		38041757	Asia
1/13/2020	13	1	2020	0	0	Afghanistan	AF	AFG		38041757	Asia
1/14/2020	14	1	2020	0	0	Afghanistan	AF	AFG		38041757	Asia
1/15/2020	15	1	2020	0	0	Afghanistan	AF	AFG		38041757	Asia
1/16/2020	16	1	2020	0	0	Afghanistan	AF	AFG		38041757	Asia
1/17/2020	17	1	2020	0	0	Afghanistan	AF	AFG		38041757	Asia
1/18/2020	18	1	2020	0	0	Afghanistan	AF	AFG		38041757	Asia
1/19/2020	19	1	2020	0	0	Afghanistan	AF	AFG		38041757	Asia
1/20/2020	20	1	2020	0	0	Afghanistan	AF	AFG		38041757	Asia

# Visualization: Bar Chart

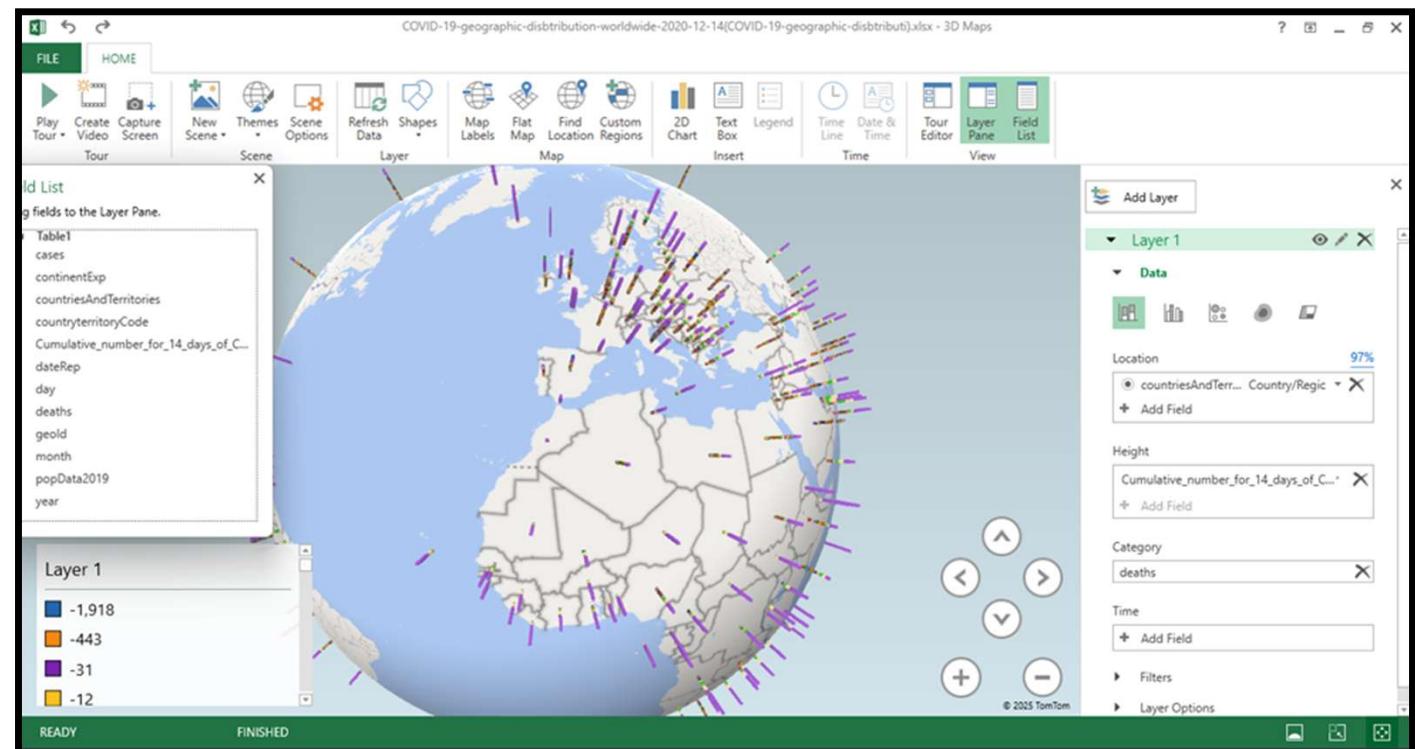


# Worldwide map of confirmed COVID-19 Cases



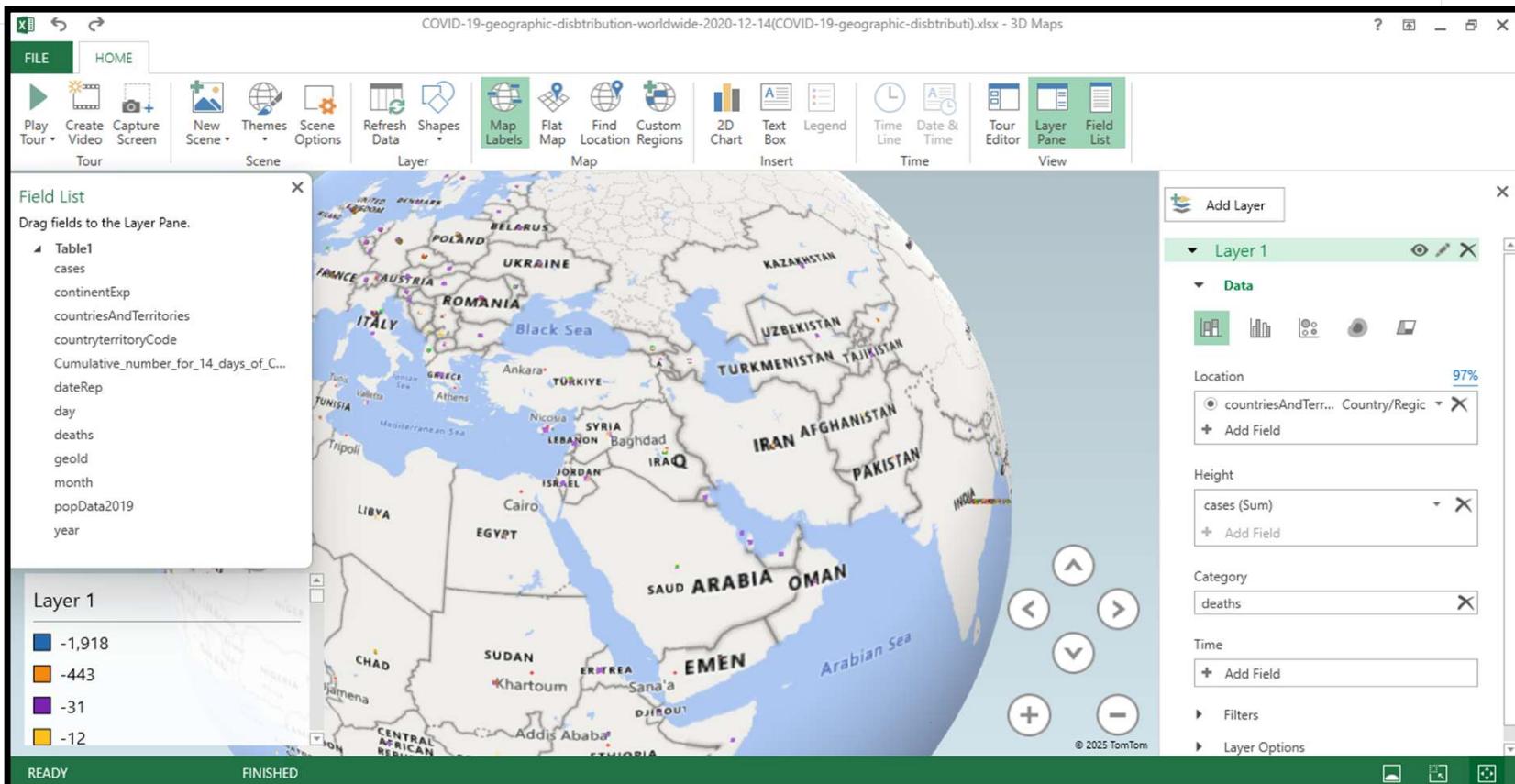
# Visualization: Excel 3D Map

- Countries and Territories
- Cumulative number for 14 days per Covid-19 cases per 100,000



# Visualization: Excel 3D Map

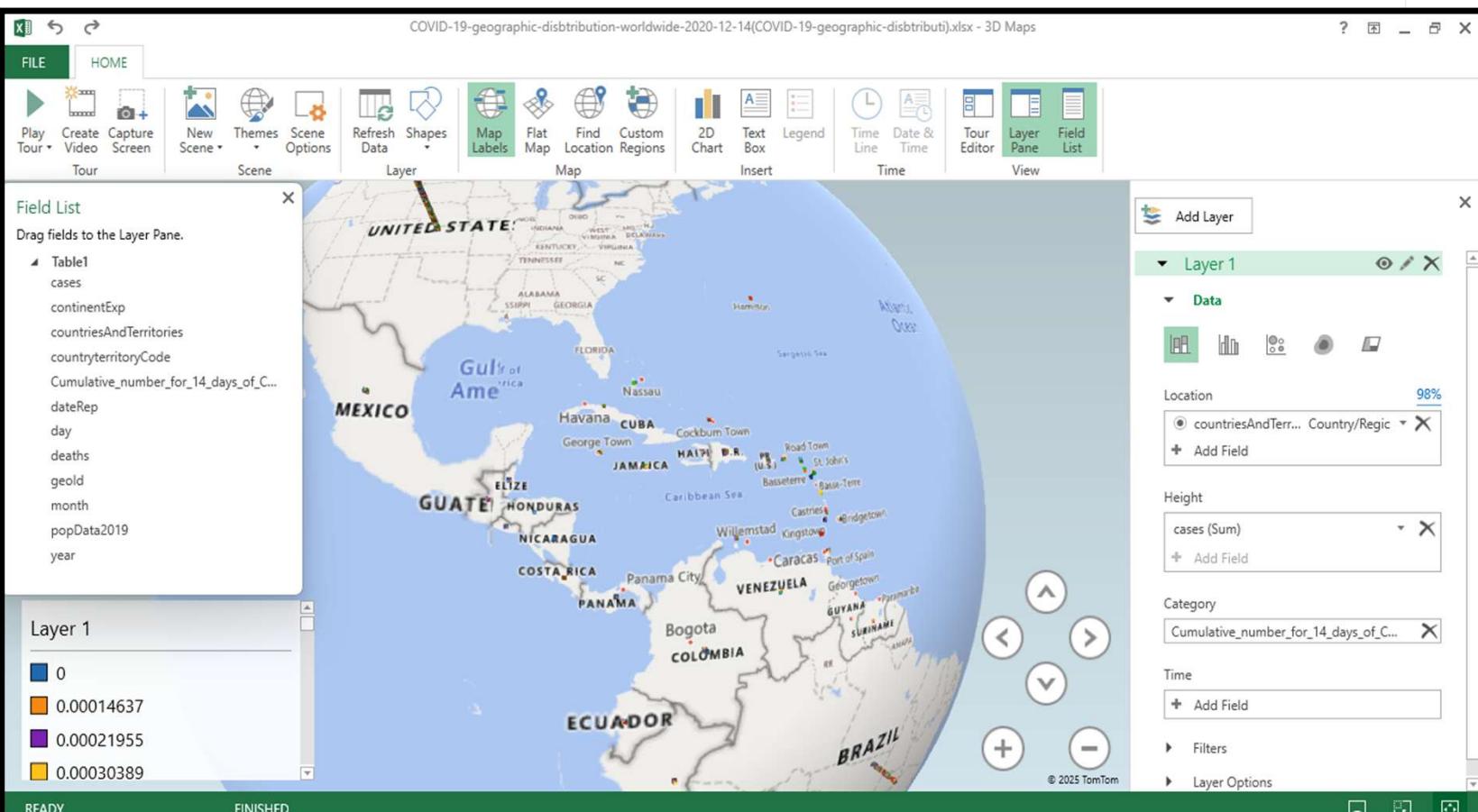
- Countries and Territories
- Cases
- Deaths



12/14/2025

# Visualization: Excel 3D Map

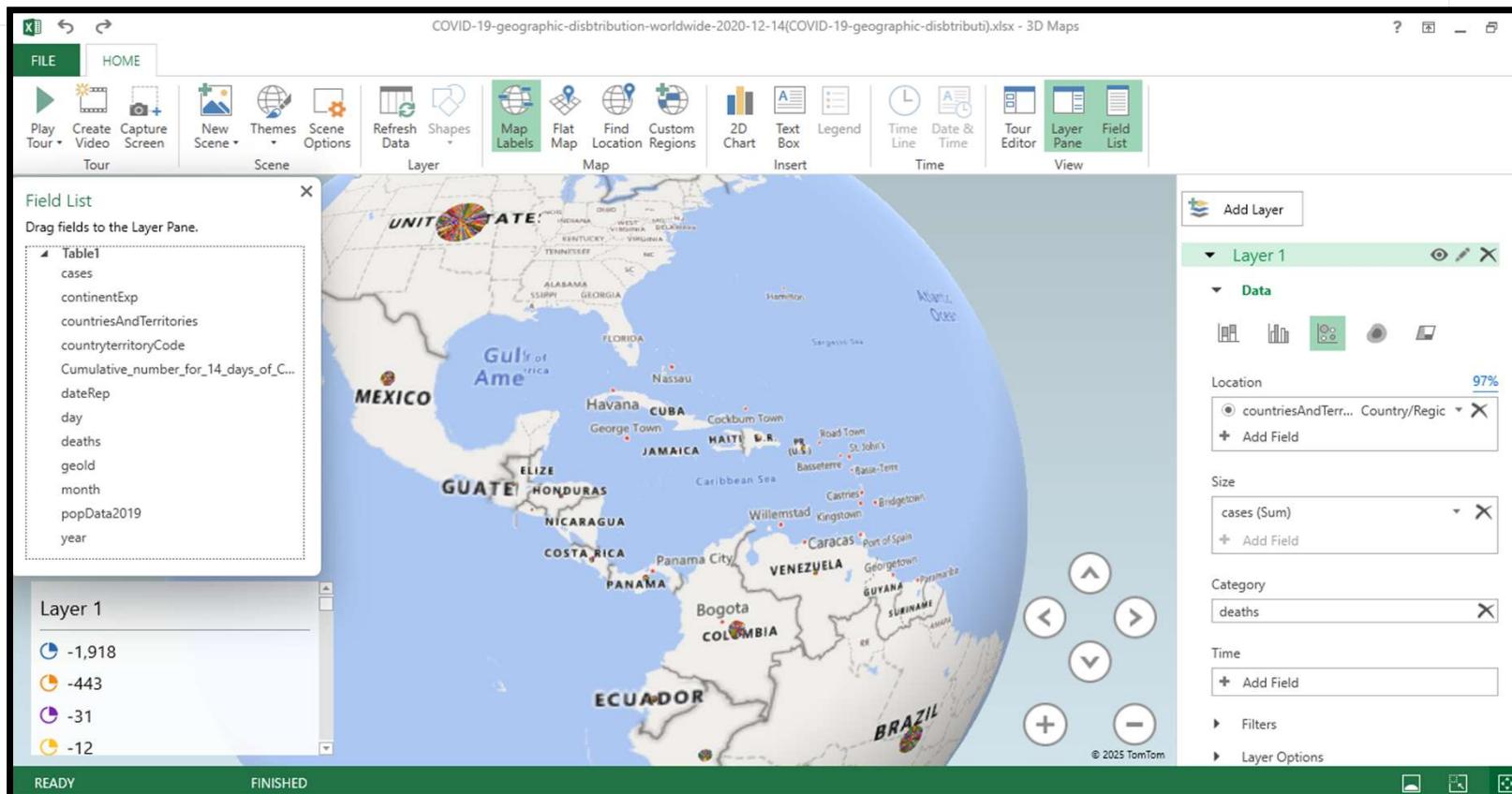
- Countries and Territories
- Cases
- Cumulative cases



# Visualization: Excel 3D Map (Bubble)

- Countries and territories
- Cases
- Deaths

12/14/2025



# Key Findings

- COVID-19 showed steep growth trends during major variants
- Countries differed significantly in peak periods and fatality rates
- Higher population density equals higher confirmed case totals
- Earlier lockdowns flattened trends for some regions
- Data confirms seasonal and global wave patterns



Thank you

Any Questions?

---