Covid Impact Analysis Using Hadoop and Apache Hive

Authors: Bryan Chica, Mohammed Shodimu, Phillip Navarro, Suleima Gonzalez
Department of Information Systems, California State University Los Angeles
CIS 4560-01 Big Data
bchica@calstatela.edu, mshodim@calstatela.edu, pnavar30@calstatela.edu,
sgonz307@calstatela.edu,

Abstract: This project uses a 20 GB+ dataset of Covid 19 records worldwide to perform a Temporal-spatial analysis to uncover the rate of spread and affected demographic of the Covid 19 infection worldwide. We analyze the data to identify hot spots and map the flow and rate of infection between different countries and detect temporal trends such as travel patterns. Insights gained from this analysis are crucial in preventing the spread of infection and for efficient distribution of funds, aid and resources. The project's visualizations will present the discovered insights, highlighting the power of big data analytics in addressing a response to a world-wide pandemic.

1. Introduction

The analysis will involve analytics of covid 19 infections per country, reported cases per demographic, population and geographic coordinates, using Hadoop and Hive. The dataset consists of a wide variety of data from around the world ranging from population (by country), population (by varying age groups), gender affected, new confirmed cases, new deceased, vaccinated and etc.., all according to reported cases from individual countries.

This particular project was chosen due to the recurring impact pandemics have on our world and societies. Though ample data exists regarding Covid, past research shows that pandemics are still able to impact communities and countries without much resistance. This is most likely due to the fact that most information retrieved from data is gathered, analyzed, and used in the exact same way without considering a much broader view of the issue. Being that this project is a Big Data project, what better reason is there that to take a step back and allow the data speak for itself and give us insights as to how pandemics work on a larger scale. As such, the dataset was gathered with that issue in mind whereby every individual dataset will be compiled into cumulative sets to give a broad picture of where, how and why pandemics spread and affect lives.[1] With the approval of CSULA faculty member, Dr. Jongwook Woo, we were able to retrieve a 20 GB+ dataset from Google Cloud.

Unlike many existing COVID-19 studies that focus on either highly granular spatial analysis or AI based detection, this project emphasizes cumulative global analysis usingdistributedd big data technologies. By leveraging Hadoop and Apache Hive in a cloud environment, our approach focuses on scalable data aggregation, efficient querying, and reproducibility across a large dataset exceeding 20Gb. This framing allows the data to be analyzed at a global level while maintaining a clear and repeatable workflow suitable for large scale analytics.

2. Related Work

Since Covid isn't a new phenomenon, a large number of studies and analytics have been performed for it using publicly accessible data. Some of the more accessible data include free data on online repositories such as Kaggle and Google Cloud.[2]

An equally valid project performed on Covid one that creates architecture for Covid-19 analysis and detection using big data, AI and data architecture. The work is available as a doc on National Library of Medicine: National Center for Biotechnology Information (August 2024) and was performed by Ahmed Mohammed Alghamdi 1, *, Waleed A Al Shehri, Jameel Almalki, Najlaa Jannah, Faisal S Alsubaei. This analysis offers valuable insight on how AI architecture can help track and detect patterns in the Covid outbreak via the use of blockchains. It does this by creating a data pipeline; leading from the data acquisition at the Healthcare, through the Big Data and Analysis using AI by the Technical Administrators and down to the Stakeholders. One major issue they face includes the sheer amount of data being too much for the available resources. Diagrams displayed the architecture of the proposed system.

A second project worth noting is the Covid-19 Open-data: a global-scale spatially granular dataset for coronavirus by Oscar Wahltinez, Aurora Cheung, Ruth Alcantara, Donny Cheung, Mayank Daswani, Anthony Erlinger, Matt Lee, Pranali Yawalkar, Paula Le', Ofir Picazo Navarro, Michael P. Brenner and Kevin Murphy. This work has many similarities to our project in that though it gathers less dataset, it still uses it to analyze the impact and patterns of Covid with regards to Population, demographics i.e. sex, age, vaccinations, etc., They take in datasets from over 200 countries and compares it with data from establishments like John Hopkins University. This work also utilizes interactive tools for visualization. However, this project differs with ours in one key area in that it implements big data on a granular scale while our project works on a cumulative scale and utilizes this analysis to provide insights on a much more global scale.

3. Specifications

The dataset is a 20.9 GB dataset retrieved from Google Public Dataset that covers data collected from countries across the globe over a span of 14 days. Due to the limitations of the project being set at 2 GB, our dataset received the appropriate approval after we showed our ability to clean the data down to 10% percent of the original data. Baring a few irregularities in the tables, we were able to insert the data into our workflow and architecture and clean the data down to specifications.

Table 1 shows files and size of the files from the dataset.

Table 1: Data Specification

Source	Google Public Dataset
File name	aggregated.csv
Extracted size	20.9 GB
Compressed size	1.32 GB
Records	3,000,599
Columns	708
Format	.csv
Data types	STRING, FLOAT, INT, BOOLEAN, DATETIME

Below, Table 2 details the System Resources and specifications for the system we will be using for this project.

Table 2: System Resources

Host name	bigdaiun0.sub03291929060.trainingvcn.oraclevcn.com
O. S.	Oracle Linux Server 7.9
CPU	AMC EPYX 7J13 64-Core Processor
Clock speed	2.45 Ghz
Storage	120 GB
Memory size	32 GB (8 GB swap)

Below, Table 3 shows the specification for the cluster in use and the Hadoop specifications.

Table 3: H/W Specification

Cluster version	Hadoop 3.1.2
Number of nodes	3

4. Architecture and Flowchart

As mentioned, the raw dataset was downloaded from a trusted Google cloud public data source. The dataset contains 708 columns of data comprising of information like from population (by country), population (by varying age groups), gender affected, new confirmed cases, new deceased, vaccinated. The flowchart below shows the data handling process (Figure 1). The dataset was accessed and uploaded via Linux to the Hadoop File System. After which it is queried using HiveQL to create the tables' schema is created. The data is cleaned, and the results are exported as a csv file. The csv output file is downloaded and uploaded to Tableau Public and Excel which are then used for visualizations.

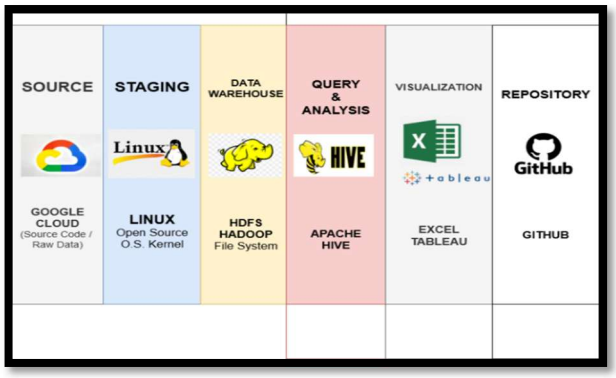


Figure 1 – Architecture

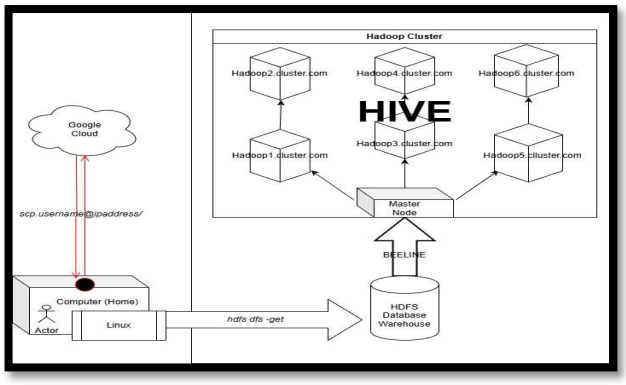


Figure 2 - Flowchart Implementation

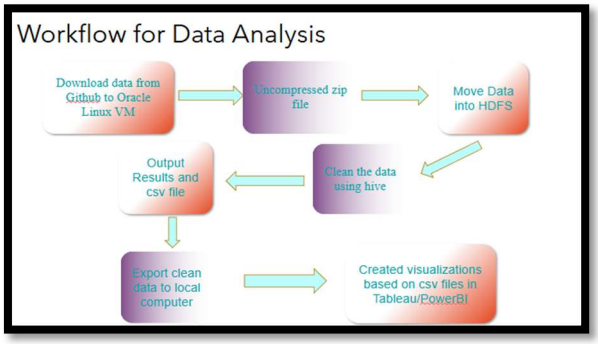


Figure 3 – Workflow of Data Analysis

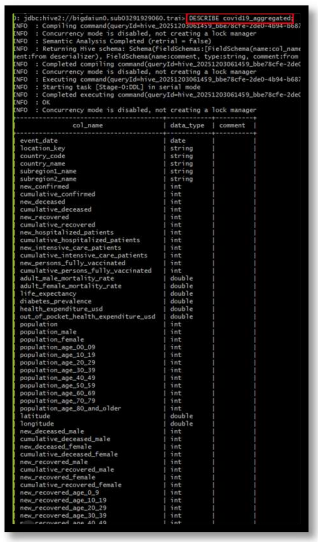
5. Data Organization

The Raw files were uploaded via Linux and stored in HDFS. We used Beeline to access Hive where we created the appropriate tables for the data. Using the tables created we loaded and organized the data into tables and then exported the data for further analysis by downloading it onto HDFS

where directories were created. Due to the massive size of the data, queries and statements were used to clean up the data into a more manageable size; roughly about 10% of the total dataset. Using specific steps, we created the tables needed to organize the data and resolve any issues:

Download data from Google Public dataset:

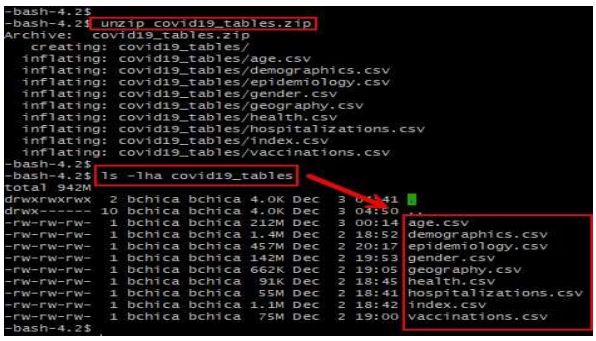
wget -O covid19_aggregated.csv.gz https://storage.googleapis.com/covid19-openhttps://storage.googleapis.com/covid19-open-data/v3/aggregated.csv.gzdata/v3/aggregated.csv.gz



Extract dataset using gunzip

gunzip covid19_aggregated.csv.gz

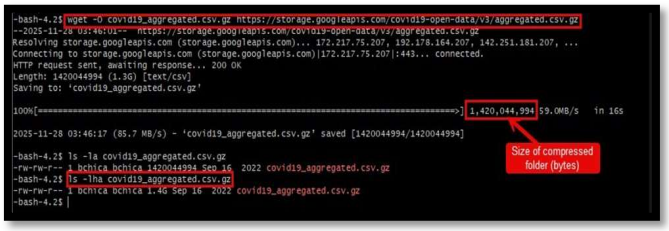
ls -lha covid19_aggregated.csv



Or--

Retrieve cleaned data from GitHub:

wget https://github.com/chica-94/COVID19-Data-Analysis/raw/refs/heads/main/covid19_tables.zip



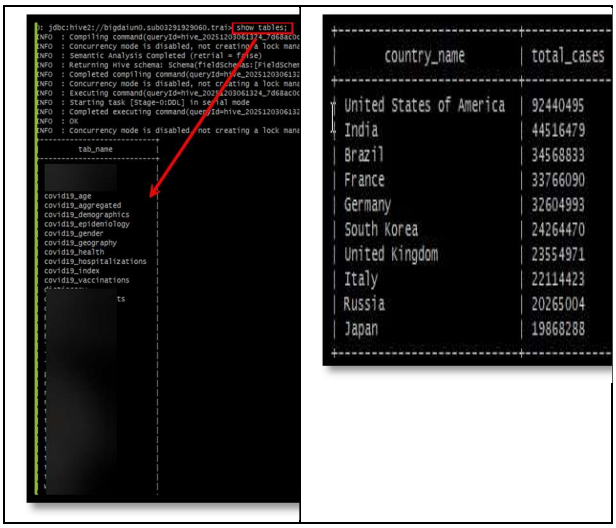
We retrieved the top 10 countries

```
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
```

Created empty table called covid19_top10_countries, which contained the cumulative confirmed cases for each country (city & states) from covid19_aggregated tables. Once the data was 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;
```



Make directories in HDFS and put data into tmp folders

```
-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$
```

6. Analysis and Visualization

Further data cleaning was made with regards to some misplaced headers for categorizing the data in the csv output file i.e., Date-time. The csv file is extracted and uploaded into Tableau Public and Excel in order to make use of Excel's 3D mapping to show Covid spikes and flow over a set time in different chart types.

6.1 Tableau Chart

The first visualization (Figure 4) is a bar chart created on Tableau Public to show the comparison between the 747,070,973 Cumulative confirmed cases versus the New confirmed cases which came in at under 1 million. The simple bar chart shows the almost negligible contribution of new cases to the overall confirmed cases, which could signify a drop in overall cases or an extremely slow build of new cases worldwide.

Figure 4 – Bar chart of Cumulative versus New cases

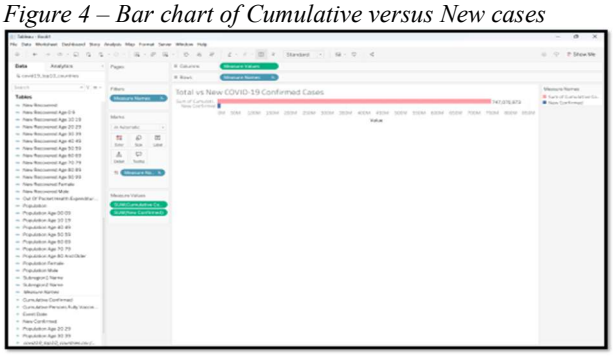


Figure 5 -Worldwide map of confirmed Covid-19 cases

6.2 Excel

Excel 3D map is used to render a 3D visualization map of both a stacked 3D map and a bubble 3D map in (Figure 3), and the location field "Countries and Territories" is used for geo-coding. Longitude and latitude are implemented to give us specific readings on the Covid cases in individual countries. The 3D map highlights the United States, India, Europe and Brazil as the highest impacted regions while Africa shows relatively low numbers.

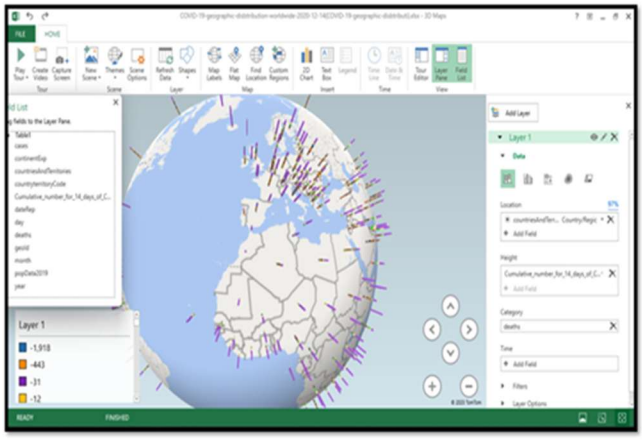


Figure 6 – Cumulative number for 14 days per Covid-19 cases per 100,000

The 3D map shown in Figure 6 shows that the Cumulative number of cases per Covid-19 cases per 100,000 people cross referenced against countries and territories for geo-mapping. The data is cumulative of specific demographics including gender (male and female), population by age groups and vaccination statuses. The layer designations include -1,918 representing the number of deceased. -443 representing the number of new confirmed cases, all spanning a total of 14 days.

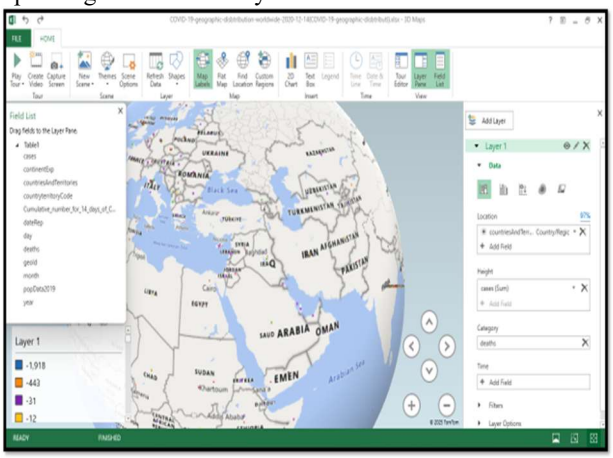


Figure 7 – Stacked showing cases per death in India and Europe

The next visualization (Figure 7) is a stacked chart with map labeling turned on for more accurate reading. This map shows the countries highlighted against the cumulated cases per 100,000 and the number of deceased, male and female, across all nine different age groups: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79 and 80-89. This includes vaccinated and unvaccinated cases.

The layer designations are the same as before: -1,918 representing the number of deceased. -443 is the number of new confirmed cases, all spanning a total of 14 days. The next maps (Figure 8) is an interactive tour which can be played to show the progression of the Cumulated Covid cases across the U.S. and Brazil in the span of 14 days.

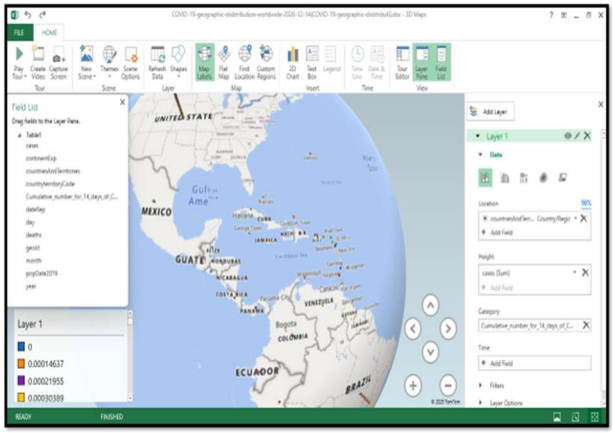


Figure 8 – Stacked chowing the culmination of cases in the United States and Brazil

6.3 Power B.I.

Additional Power B.I. visualization is a world map that highlights similar data to the excel and Tableau visualizations. This includes sum of cumulative confirmed cases and the countries involved.

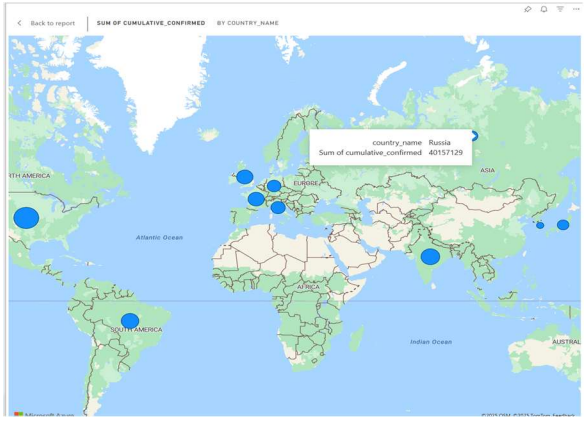


Figure 9 – Sum of Cumulative confirmed cases

7. Conclusion

Finally, from the above analysis and work we concluded the following:

i. High population density regions like India and Europe are high impact areas.
ii. Subtropical regions, i.e., Africa, are low impact regions.
iii. Regions with less vaccinations are more susceptible to transition from confirmed cases to death cases.
iv. Confirmation of previously held hypothesis of older age groups being more susceptible to the virus and thereby impacted a greater degree.

Beyond the analytical results, this project demonstrates the practical value of using Hadoop and Apache Hive to process and analyze large scale pandemic data in a distributed environment. The cumulative and global nature of the analysis, combined with a reproducible workflow, differentiates this work from smaller scale and purely visualization drive studies. This approach highlights how big technologies can be effectively applied to real world global problems while remaining accessible for future extension and reuse.

Given the available data in just fourteen days, we were able to gain valuable insight into the regions impacted higher than others and the possibilities of why. This opens the door for further discussion on what can be done in the high impact regions to mirror the quality of the low-impact regions. A more defined dataset will enable us to create and assess better data-driven analysis and solutions on the Covid-19 pandemic as well as any future pandemic.

For more information, dashboards and code visit project's GitHub link[1].

References

[1] Google Cloud, Big Query (April, 2020) https://console.cloud.google.com/bigquery?p=bigquery-public-data&d=covid19_ecdc_eu&page=dataset&project=still-algebra-475823-k6&ws=!1m4!1m3!3m2!1sbigquery-public-data!2scovid19_ecdc_eu

[2] Ahmed Mohammed Alghamdi et all., National Library of Medicine, An architecture for COVID-19 analysis and detection using big data, AI, and data architectures. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC11293665/

[3] Oscar Wahltinez et al. Scientific Data. (April 2022) Covid-19 Open-data. A Global Scale spatially granular meta-dataset for coronavirus disease. Retrieved from https://www.nature.com/articles/s41597-022-01263-z

[1] GitHub Link: https://github.com/chica 94/COVID19-Data-Analysis