**Covid-19 Dataset Analysis Using Hive**

Authors: Timothy Rochester, Jae Hoon Lee, Riker Santivong, Wesam Farjo, Justin Licea

Department of Information System, California State University Los Angeles

CIS 4560-01 Introduction to Big Data

Email: [troches@calstatela.edu](mailto:troches@calstatela.edu), [jlee464@calstatela.edu](mailto:jlee464@calstatela.edu), [rsantiv@calstatela.edu](mailto:rsantiv@calstatela.edu), [wjargee2@calstatela.edu](mailto:wjargee2@calstatela.edu), [jlicea7@calstatela.edu](mailto:jlicea7@calstatela.edu)

**Abstract:**Various institutions have allocated data on COVID-19 cases and vaccinations worldwide in publicly available datasets. These datasets not only include raw numbers also include multiple statistics such as cases per million. This report will analyze different countries' COVID-19 case rates and COVID-19 vaccination rates. We will also examine the sentiment of users on Twitter in those countries.

**1. Introduction**

The first confirmed case of COVID-19 was discovered in Wuhan, China. Since then, it has exploded into 147 million confirmed cases and over three million deaths worldwide, and it is the severe global health emergency since the 1918 influenza pandemic. As a result, there are no shortage of data and models about this pandemic. Our report will use new COVID-19 cases and vaccination rates and compare the former with the sentiment of tweets over time. The countries that will be included are the United States, Israel, India, United Kingdom, Chile, Bahrain, and Hungary. This project provides information about tweet sentiment analysis which is used to determine the feelings of Twitter users in the United States over the course of the pandemic.

**2. Related Work**

Twitter is a popular social media app that allows users to create short posts containing the text of up to 140 characters. With 192 million daily users, the number of messages sent is vast. Because of this, Twitter is a rich source of data for researchers to use for sentiment analysis. Researchers use sentiment analysis to determine the public's general opinion on a specific topic. This process is commonly done by assigning a value to terms that have been classified as "negative," "neutral," or "positive" and then giving a value to each classification. Next, an average of each classification is calculated to determine general sentiment.

The Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study published on November 11th, 2020, describes a study that also sought to gauge the public's sentiment of COVID-19. They did this by using an application programming interface (API) to collect tweet data from December 13th and March 9th, 2020 [1]. The data we used for our analysis spanned a more extended amount of time – from January 28th, 2020, to January 1st, 2021. Another way this project differed from our own is that the researchers used Python and RStudio to analyze their data, whereas our team used Hive and Excel.

Our project also looks at relationships between the rate of vaccinations and new Covid-19 cases per million in each country. The CDC's Covid Data Tracker contains maps and charts tracking cases, deaths, and other trends of COVID-19 in the United States [2]. The paper, Covid-19 Dynamics After a National Immunization Program in Israel, studied the effects of the Pfizer vaccine rollout in Israel and sought to determine the effectiveness of the vaccine deployment. The study results show a decline in hospitalizations due to Covid-19 3-4 weeks after the start of the national vaccination campaign [3]. Our project was similarly interested in the temporal changes that occur after vaccination. The main differences are our focus on multiple countries and our utilization of Hive to manage the datasets we were working with.

**3. Background**

To join tables used in our project, we referenced work done in Lab 3 [8]. We employed geo-temporal visualization based on Twitter sentiment in our data analysis. Our geo-temporal visualization is based on work done in our Lab 4 coursework [9]. By applying methods introduced in the lab, we were able to clean our Twitter data and ready it for visualization using an Excel 3D map.

**4. Specification**

We used three datasets in CSV file format for this project. The datasets consisted of global coronavirus cases and vaccination data and Twitter tweet data from the United States. The total size of these files is 5.58 GB, which consists of Covid19\_case.csv, tweetid\_sentiment.csv, and world\_ vaccination.csv. The new datasets of COVID-19 cases and world\_vaccinations cover all countries in the world.

However, the tweet sentiment dataset only has data over the United States.

*Table 1 Specifications of SCU Hadoop Server*

|  |  |
| --- | --- |
| Version | Hadoop 3.1.4 |
| Total Nodes | 3 (x 16 cores) |
| Total Node Memory Size | 192 GB |
| RAID | 24TB/18TB (RAID) |
| PySpark | Python 2.7.5, Spark 2.3.2.3.1.4.0-315 |

**5. Our Work**

**Data Sources**

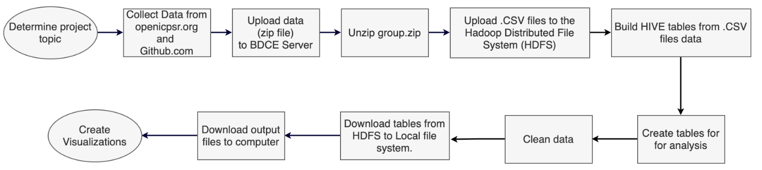
Covid-19 World Case Dataset 19.3 MB [4]

Covid-19 World Vaccination Dataset 847 KB [5]

Global Reactions to Covid-19 on Twitter: A Labelled Dataset with Latent Topic, Sentiment and Emotion Attributes 5.56 GB [6]

State-locations file [7]

**Preparing Data**



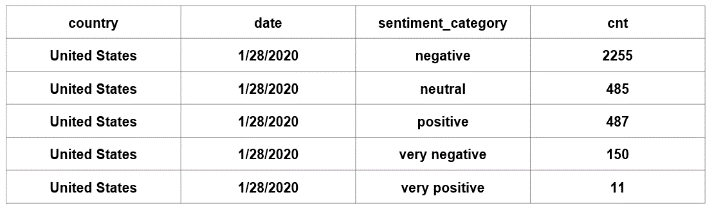
*Figure1 - Project Workflow*

**6. Analysis 1: Twitter Sentiment**

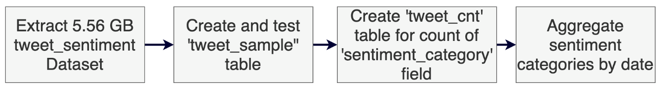
The first analysis our team did was analysis on tweet sentiment about the COVID-19 pandemic during 2020. Our team used a premade dataset containing tweets about the pandemic sorted into sentiment categories labeled by country of origin, date, and number of tweets. The goal of the analysis is to evaluate the sentiment of Twitter users in the United States on Covid-19 over time.

Data: tweet\_cnt.csv:

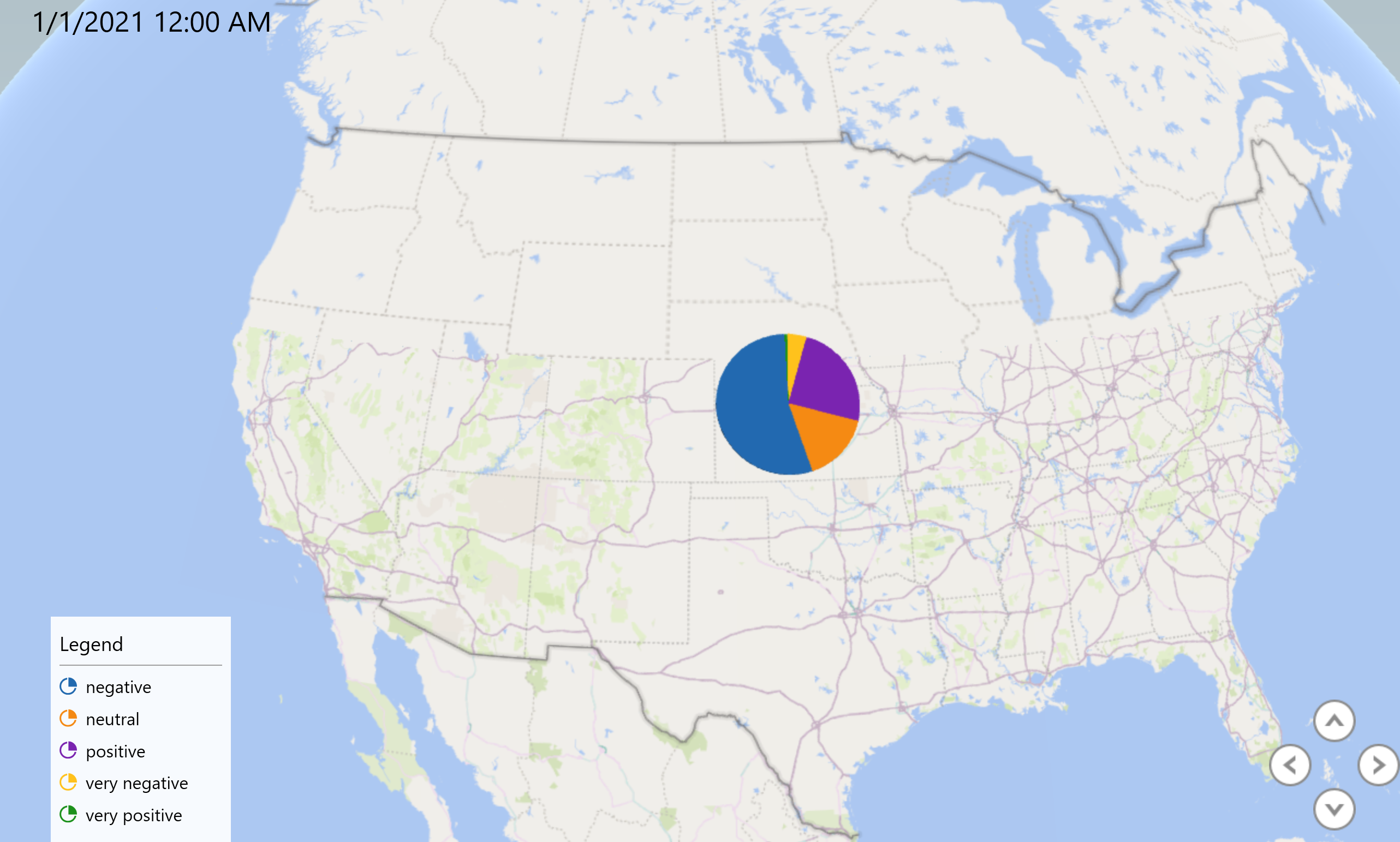
* Total Rows: 1,707​
* File Size: 71.6KB​
* Computation Time: 67.469 seconds​

 *Table 2 tweetid\_sentiment*

To create this analysis, we utilize Hadoop (HDFS), HIVE query language, and Excel-3D Maps.

First, we extract data from the ‘tweetid\_sentiment’ file to create the table. Then we created the 'tweet\_sample' for a small table for testing. Finally, we created the 'tweet\_cnt' table from ‘tweetid\_sentiment’ table to count the 'sentiment\_category' fields. 

*Figure 2: Workflow: Twitter Sentiment Analysis*



*Figure 3 - Geo-temporal visual using 3D mapping*

Overall sentiment breakdown in the United States from 1/28/2020 – 1/1/2021 is:

54% Negative,

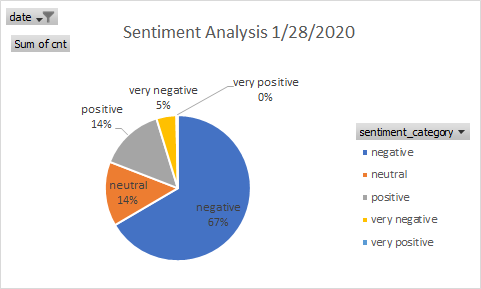
6% Neutral

25% Positive

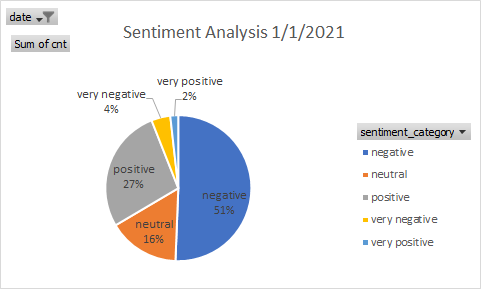
4% very negative

1% very positive

There is an increase in positive sentiment and a decrease in negative sentiment over the year.



*Figure 4 – Twitter Sentiment 1/28/2020*



*Figure 5 - Twitter Sentiment 1/1/2021*

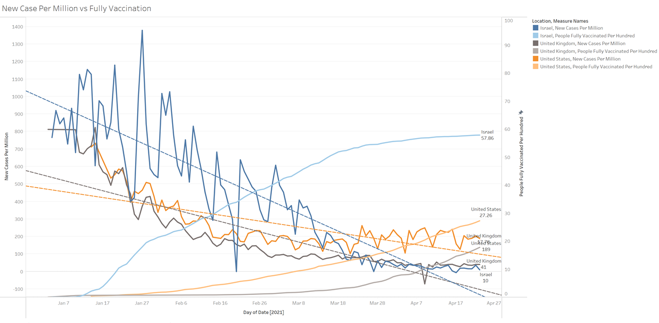
**7. Analysis 2: Vaccinations & New Cases**

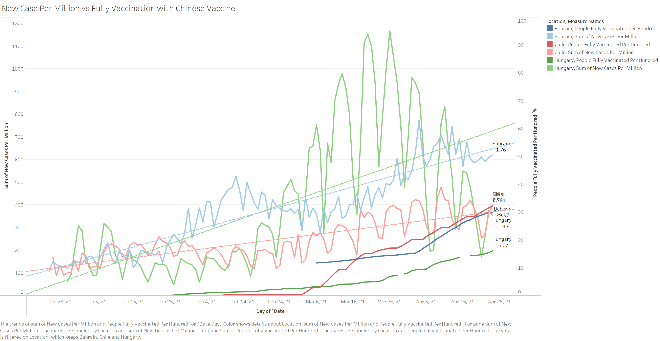
Our second analysis was on the relationship between vaccinations and the number of new cases of COVID-19. Beginning in early 2021, countries began to vaccinate their populations, with those in the west generally doing the most. We measured new cases per million and the number of fully vaccinated people per hundred people between January 1st and April 23rd in certain countries in the world.

Data: vac\_case.csv

* Total Rows: 14,110
* File Size: 1.03MB
* Computation Time: 14.346 seconds

**Workflow:** First, we extracted the data from the COVID-19 cases and the vaccination datasets. To prepare sample tables, we need to change the date format from string to data for joining. Then we organized and joined two tables (covid19\_case\_sample1 and world\_vac\_sample1) in Hadoop and Hive. Finally, we loaded the data into Tableau and used it to create two line charts. The darker colored line represents new COVID-19 cases per million, and the lighter colored line represents fully vaccinated people per hundred. The dashed line is a trend line for COVID-19 new cases per million.

 *Figure 6 - Workflow: Vaccinations & New Cases*  *Figure 7 - Trends in Israel, United Kingdom, and United* States

*Figure 8 - Trends in countries (Bahrain, Chile and Hungray) that mainly use the Sinopharm vaccine.*

Key findings are:

        I.Generally, in countries that vaccinated a notable number of their population, as vaccinations increase, the number of new cases decreases.

      II.In countries that mainly use the Sinopharm vaccine [4], new cases are increasing despite increasing vaccination rates.

a. Chile is the best example as 94% of its vaccines come from China [11].

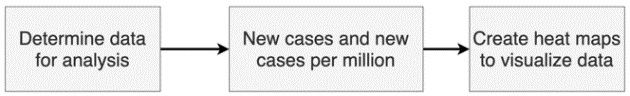
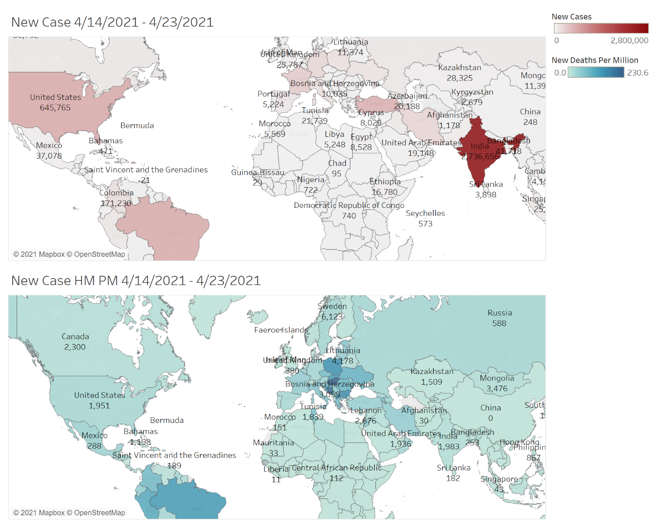
**8. Analysis 3: COVID-19 New Cases**

The third analysis was on the number of new cases in a ten-day period around the world. Our team measured cases from April 14th to April 24th throughout the world and displayed them on a global heatmap. We analyzed the new case in two different perspectives to display the data: new cases and new cases per million. The goal was to analyze the new cases of COVID-19 in the world by these two measures.

Data: Covid19\_case\_sampl1.csv

* Total Rows: 833,868
* File Size: 5,926 KB
* Computation Time: 16.45 seconds

Using the covid19\_case\_sample table, which was extracted and cleaned in Analysis 2, our team determined the data used for the analysis. We selected the new cases and new cases per million and then used that data to create a global heatmap. The severity of the outbreak scales the color according to the metric; the darker the color, the worse. The analysis zoomed in to focus on certain regions of the world, mainly the Northern Hemisphere.

*Figure 9 - Workflow for New Cases per Million*  *Figure 10 - Heatmap of new cases and new cases per million*

Key findings are:

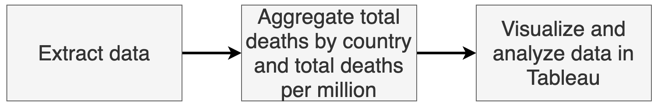
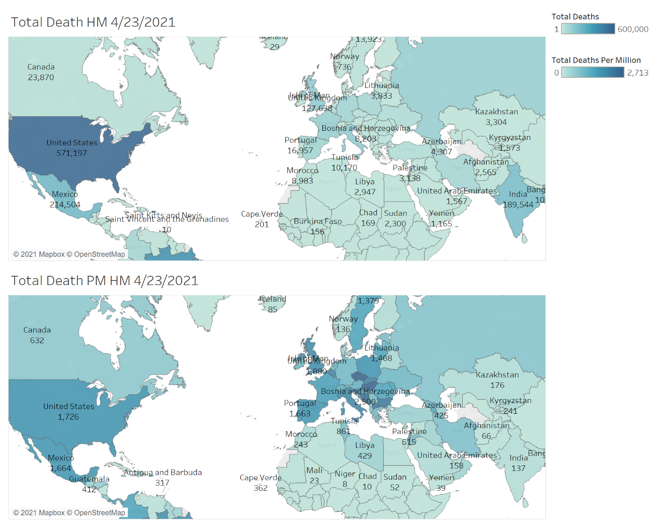
      I.There are significant differences in perspective between total new cases and new cases per million.

    II.India has the highest number of new cases, but new cases per million showed that many countries have more severe outbreaks compared to their population.

**9. Analysis 4: Total Deaths**

The last analysis we did was all deaths caused by COVID-19 worldwide. Our team measured all counted deaths caused by COVID-19 up to April 23th 2021. Like Analysis 3, our team used two metrics: total death and total death per million. The goal was to analyze the total deaths of COVID-19 in various countries worldwide by these two measures.

We used a similar procedure as we did in Analysis 3. Using the data extracted and cleaned in Analysis 2, our team determined the data used for the analysis. We selected the total deaths and total deaths per million and then used that data to create a global heatmap. The sum of deaths scales the color according to the metric; the darker the color, the worse. The analysis zoomed in to focus on certain regions of the world, mainly the Northern Hemisphere.

 *Figure 11 - Workflow for total deaths*  *Figure 12 - Total deaths heatmap*

Key Findings are:

      I.Like analysis #3, there is a notable discrepancy between the heatmaps created by the two measures.

    II.The total death showed the United States is the highest with 571,197 deaths. However, Hungary is the highest with 2,713 total deaths per million.

**10. Conclusion**

Finally, summing up all the above work, we can conclude the following:

1. Still, the majority of Twitter sentiments are negative.
2. Depending on the brand of the vaccine, it will affect the suppression of new cases.
3. New cases between 4/14/2021 and 4/23/2021 are highest in India but highest in per million Cyprus.
4. Total deaths in 4/23/2021, the United States has the highest number of deaths, but per million Hungary has the highest.

By utilizing Hive and the Hadoop Distributed File System, we managed, prepared, and processed large amounts of data to create data visualizations from the raw data files. The visualizations we created allowed us to identify relationships between vaccination rates and decreases in new cases. They also made us question the efficacy of vaccines distributed in nations with vaccination rates and still climbing new cases.

**References**

[1] Boon-Itt, S., & Skunkan, Y. (2020). Public

perception of the covid-19 pandemic on twitter:

Sentiment analysis and topic modeling study. *JMIR*

*Public Health and Surveillance,* *6*(4).

doi:10.2196/21978

[2] CDC COVID Data Tracker. (n.d.). Retrieved May 13, 2021 from <https://covid.cdc.gov/covid-data-tracker/#/datatracker-home>

[3] *China's Vaccine Map: Countries using Chinese*

*vaccines*. 404. (2021, March 14).

<https://news.cgtn.com/news/2021-03-14/China-continues-to-contribute-to-global-vaccine>

[4] Covid-19 World Case Dataset (2021). Retrieved from

<https://raw.githubusercontent.com/owid/covid-19-data/master/public/data/owid-covid-data.csv>

[5] Covid-19 World Vaccination Dataset (2021). Retrieved from <https://raw.githubusercontent.com/owid/covid-19-data/master/public/data/vaccinations/vaccinations.csv>

[6] Global Reactions to Covid-19 on Twitter: A Labelled Dataset with Latent Topic, Sentiment, and Emotion Attributes (2021). Retrieved from

<https://www.openicpsr.org/openicpsr/project/120321/version/V6/view%3Bjsessionid=947AC84E86D639FAE3F8F846446C91B0>

[7] J. Lee, “The State Location File.” (2021). Retrieved from <https://drive.google.com/file/d/1GDLgobvZK_tVDiZgF3otgzyOtIkvWZAA/view>

[8] J. Woo, “Lab: HVAC Sensor Data Analysis in

Hive using Oracle Cloud,” Dept. Inf. Syst.

California State University Los Angeles, CA, USA,

2020

[9] J. Woo, “Lab: Hive Twitter Sentiment Analysis

using Oracle Cloud,” Dept. Inf. Syst.

California State University Los Angeles, CA, USA,

2020

distribution-YCquwiznFK/index.html.

[10] J.lee “Covid19 case, vaccination and sentiment” from <https://github.com/abcbbong/Covid19_case_vaccination_and_sentiment>

[11] Meredith, S. (2021, April 21). Chile has one of

the world’s best vaccination rates. Covid is

surging there anyway. Retrieved

from <https://www.cnbc.com/2021/04/19/covid-chiles>

-coronavirus-cases-hit-record-levels-despite-vaccine-rollout.html

[12] Rossman, H., Shilo, S., Meir, T., Gorfine, M.,

Shalit, U., & Segal, E. (2021). COVID-19

dynamics after a National Immunization program in

Israel. *Nature Medicine*.

doi:10.1038/s41591-021-01337-2