INFO-I535 : Management Access Using Big Data

Big Data Concepts and Implementations

Report **Rahul G.S.**

Introduction

2020 saw a rapid surge in a then relatively unknown disease ‘Covid-19’. The entire world was affected in proportions never seen before. This project aims at highlighting the increase of Covid-19 cases across the ‘Monroe County (IN)’ and visualizing the trends that have been discovered through the period January 2020-November 2022.

The dataset chosen comprises of data taken though-out the entire nation. Each observation comprises of the total number of people afflicted with the disease in a specific county, with an increasing count with each column representing a passing day. The final project focuses primarily on the ‘Monroe County (IN)’, depicting and signifying spans of importance over three years.

Background

Covid-19 has been globally declared as a pandemic and is still researched upon extensively despite a drastic reduce in the number of cases. While there have been vaccines and other prevention measures to provide a safeguard against it, there is no concrete solution or cure present at the time of this report. The project hopes to provide clear timelines of importance to be further used for research and correlating the increase or decrease of any specific component responsible with the periods of prominent boom.

The data being used for visualization also holds a personal importance since it hones in on the ‘Monroe County’ present in Indiana, USA. It is a place of value because it is home to Indiana University Bloomington and my current place of residence. It was thought to be an interesting avenue to explore the changes seen before and after the commencement of the 2021-Fall Master’s Program.

Methodology

The dataset in use is chosen from the public log of ‘Big Query’ datasets available in the Google Cloud Platform. The said dataset is the Dataflix COVID dataset, which comprises of total cases of COVID-19 encountered across numerous countries. In particular, the data file selected contains the total cases in each county of the United States of America.

‘Big Query’, a database which can effectively analyze datasets in GCP is used to query the dataset. The query results in a table which is then stored in Google Drive to be used for visualization. To accomplish that, ‘Dataproc’ is used. It is a service that allows the user to take advantage of services like Hadoop, Spark or Jupyter on a cloud platform such as GCP.

Dataproc is utilized to create clusters and run a python notebook on the cloud. A bucket is established to store the notebooks and used it in the future. The notebook is then edited to perform a detailed visualization of the trends observed.

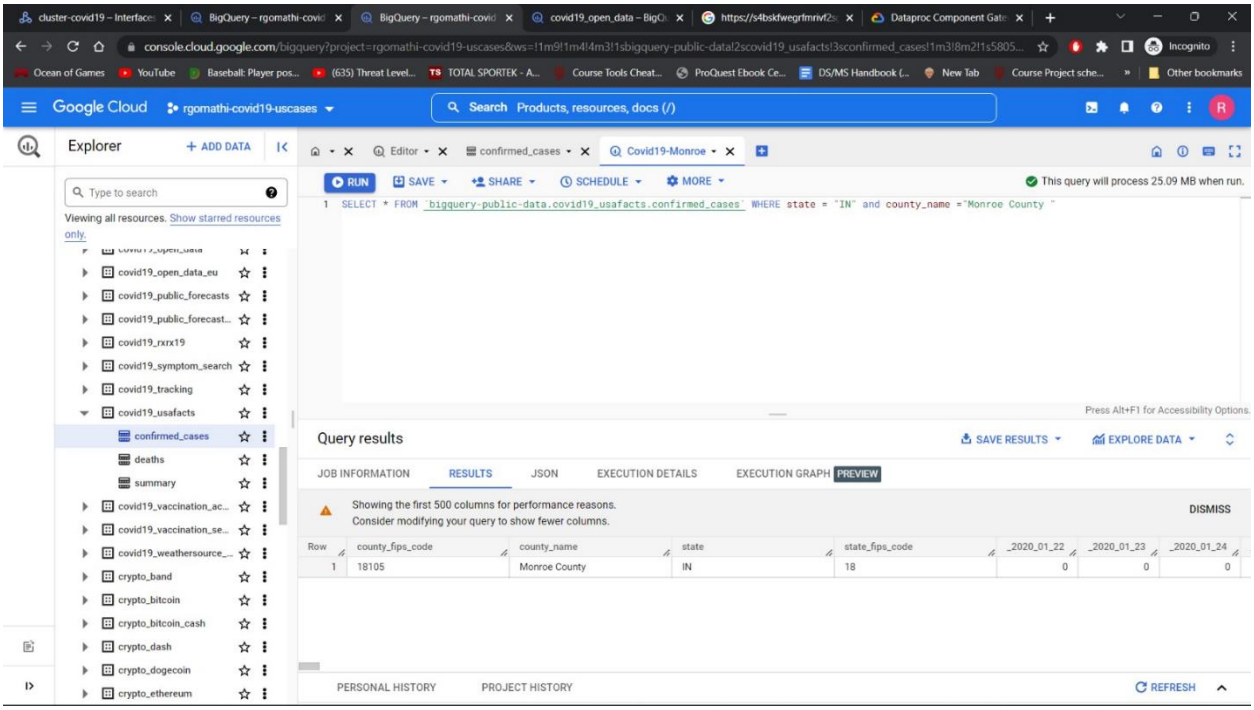


Figure 1: Dataset

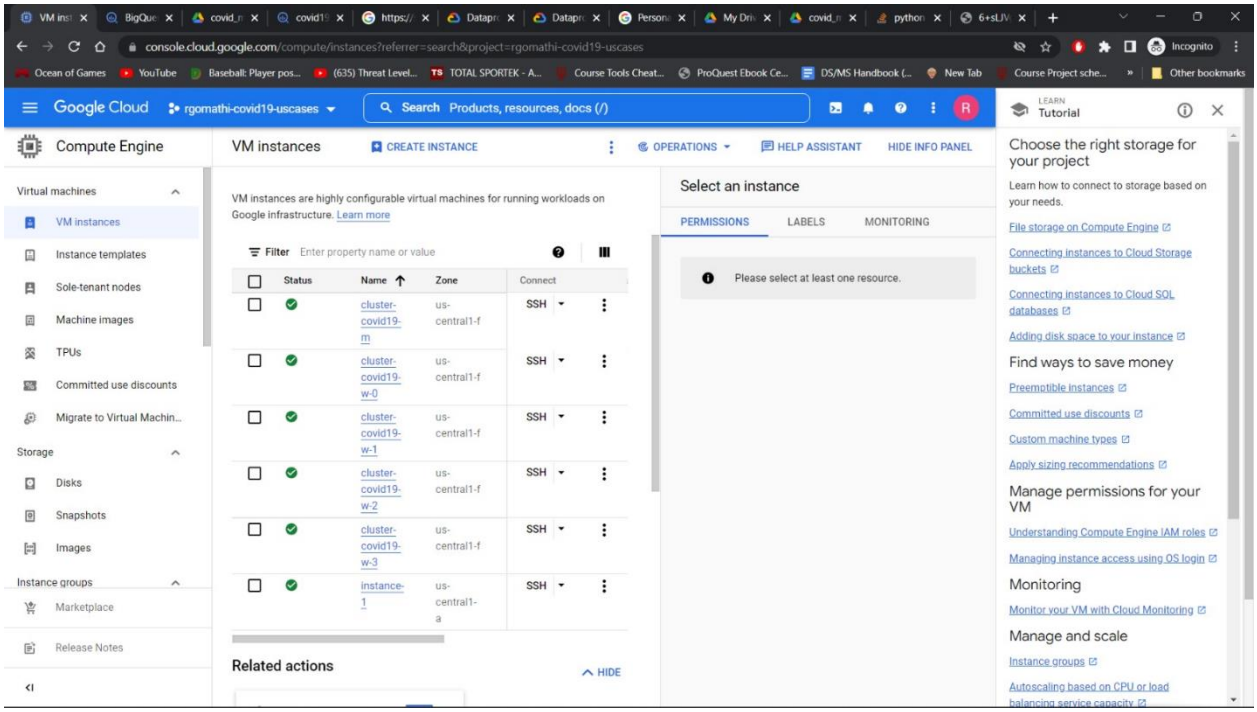


Figure 2: Clusters

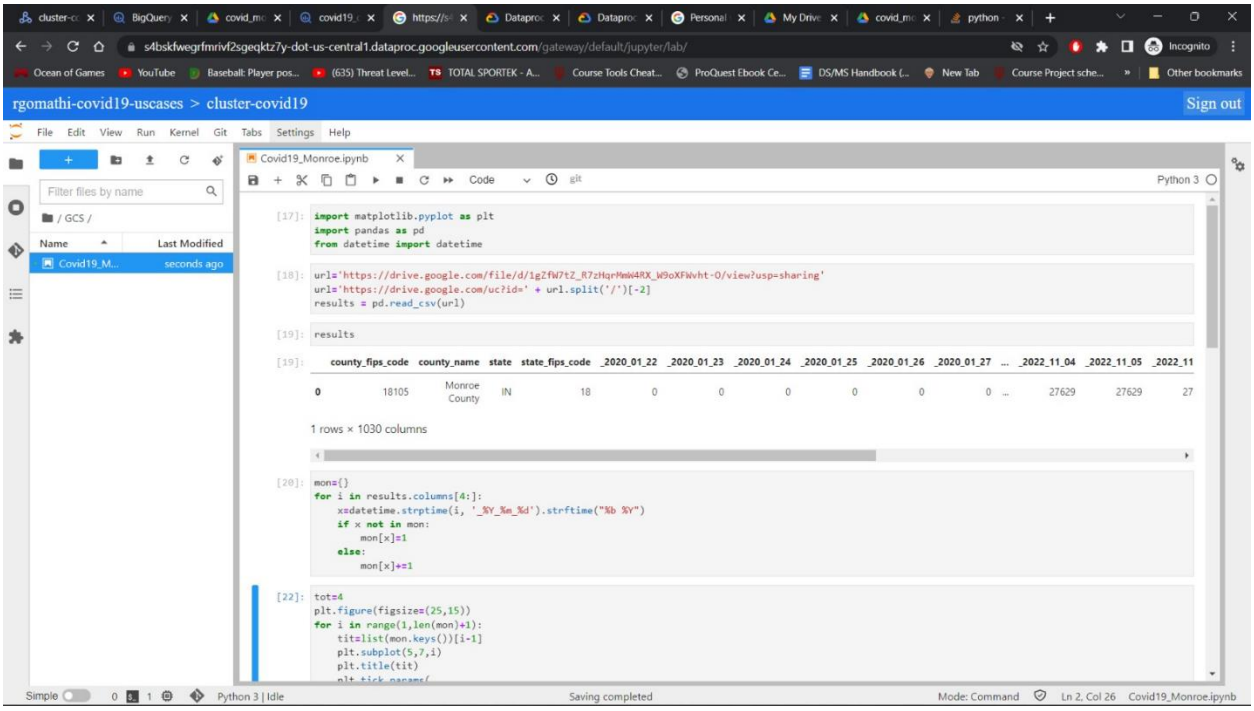


Figure 3: Jupyter Notebook

Results

The results obtained after the visualization depicts the trends of increase in total number of COVID-19 cases in Monroe County over a 35-month span. Upon closer observation, there are many underlying patterns which can be uncovered.

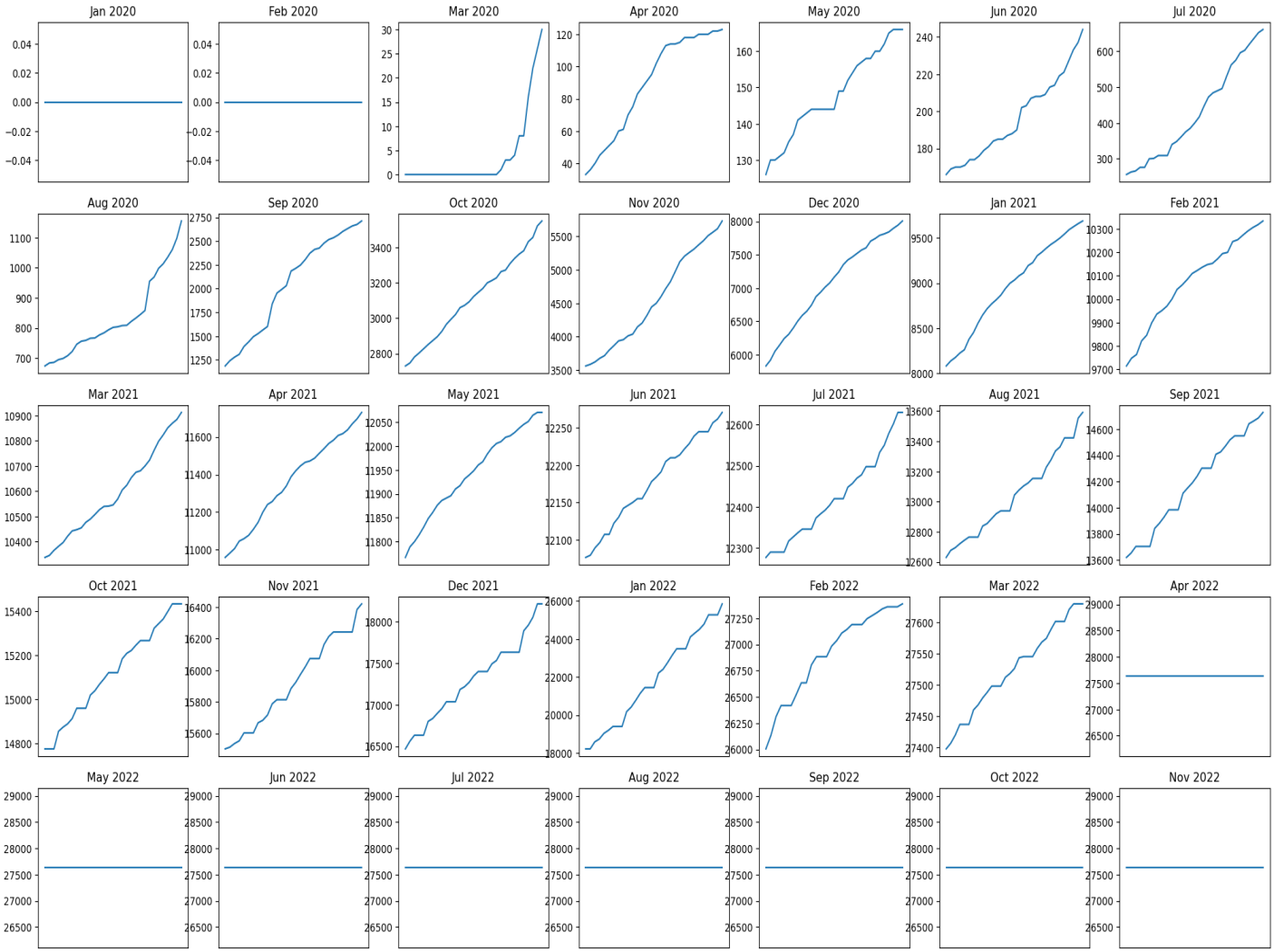
The first increase in the total number of cases can be seen in the March of 2020. This fact directly correlates with the initial spread of the disease in the said period. March 2020 saw the introduction of the disease into mainstream public across the world, and it is the same case with Monroe County as well.

The next significant upward trend can be noticed during the month of September 2020, when the general increase in cases rises from a modest 400 to about 1500, more than doubling the total number of cases in a single month. This specific trend continues through January 2021, during which the total number increases from 1100 to 9500.

The subsequent few months see a drop in the rate of increase, depicting the rise of Delta cases, a variant that afflicted fewer members than the original virus. The period between February 2021 to November 2021 underwent a relatively moderate increase, rising from 9700 to 16400.

December 2021 earmarks the uprisal of the Omicron variant, resulting in a substantial growth in rate of increase of total cases. Within the short duration between December 2021 and February 2022, the number expanded from 16500 to a whopping 27250, meaning approximately 11000 new cases at the same timeframe.

Since March 2022, the number of new cases has almost died down, with the months April to November seeing very little to no such afflicted victims. It represents the dormant nature of the COVID-19 virus in recent months.


Jan 2020
Feb 2020
Mar 2020
Apr 2020
May 2020
Jun 2020
Jul 2020
Aug 2020
Sep 2020
Oct 2020
Nov 2020
Dec 2020
Jan 2021
Feb 2021
Mar 2021
Apr 2021
May 2021
Jun 2021
Jul 2021
Aug 2021
Sep 2021
Oct 2021
Nov 2021
Dec 2021
Jan 2022
Feb 2022
Mar 2022
Apr 2022
May 2022
Jun 2022
Jul 2022
Aug 2022
Sep 2022
Oct 2022
Nov 2022

Discussion

The interpreted results show a clear set of patterns, forming a trail of how the total number of affected cases were impacted by the corresponding difference in variants within Monroe County. Visualizations of the obtained dataset depict that Omicron is the variant that affected people the most, while the Delta variant affected people in a lower level when compared to the original and Omicron variants.

The first cloud platform used in the project is Big Query. Big Query was introduced in the module ‘Complete - MongoDB and BigQuery’. The database enabled the querying of vast public datasets easily, with low computational power and speed required. The resultant tables from the query led to further research into the topic.

The other technology used in the project which was briefly touched upon in the course is Dataproc. It was given a short preface in the module ‘Complete - Distributed computing with Dataproc’. Cloud Dataproc provided an easy method of creating clusters and running subsequent jobs, aiding the usage of Jupyter notebooks in GCP. It was used to deploy distributed computing to display the various visualizations obtained via EDA and plotting.

One barrier while completing the course project was the creation of clusters. The problem ‘Global/Network/Default was not found’ kept getting encountered, something which wasn’t the case when executing the Qwiklab course of Dataproc. It was later found that the VCP network for the course project did not contain a default network, whereas the one present in the placeholder project of the qwiklab console had one. The default network was then replicated in the course project to create the clusters, but the reason behind its absence originally is still unknown.

Another difficulty was with reading the csv file from Google drive while creating the visualizations. Subsequently surfing for relevant information provided insight into splitting the URL in a specific way which makes for the dataset to be read smoothly.

Conclusion

The project conveys a clear and distinct picture of how people were affected by the different variants of COVID-19 from January 2020 to November 2022. Focusing mainly on the Monroe County of Indiana, the project makes use of Big Data concepts ‘Big Query’ and ‘Dataproc’ along with other cloud features like clusters and buckets to produce trends which can further be used in the future.

References

https://cloud.google.com/bigquery/docs/quickstarts/query-public-dataset-console#open_a_public_dataset

https://docs.bridgecrew.io/docs/bc_gcp_networking_7

https://stackoverflow.com/questions/56611698/pandas-how-to-read-csv-file-from-google-drive-public

https://www.cloudskillsboost.google/focuses/3692?catalog_rank=%7B%22rank%22%3A5%2C%22num_filters%22%3A0%2C%22has_search%22%3Atrue%7D&parent=catalog&search_id=14163071

https://cloud.google.com/dataproc/docs/tutorials/jupyter-notebook

https://www.cloudskillsboost.google/focuses/586?catalog_rank=%7B%22rank%22%3A7%2C%22num_filters%22%3A1%2C%22has_search%22%3Atrue%7D&parent=catalog&search_id=9028616

https://cloud.google.com/bigquery/public-data/

https://console.cloud.google.com/marketplace/product/dataflix-public-datasets/covid