PROJECT REPORT

INTRODUCTION:

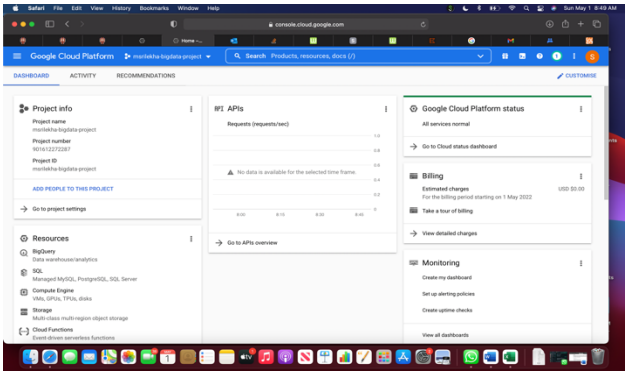
For Big Data Hands-On project, I have implemented a pipeline which consists of extract/ download, transform, and visualize. I have applied knowledge and skills learned in this course work and enhanced learning from online content available and implemented in this project. The environment I will be using in the project is Google Cloud Platform. Additionality I will be using Data Cleaning / quality assurance techniques and pipelines in this project. The motive of this project is to utilize the tools and concepts learnt from the course work, hence I will be developing a pipeline where I thought of choosing the dataset by downloading manually from the source and uploading in GCP cloud storage bucket, but as this process is manual, to automate the pipeline refresh process I will select the public datasets available in GCP marketplace. I will then perform transformations on the data and create the required insights and store in a BigQuery table. I will schedule the query with relevant frequency, so that dataset refresh and transformation steps will be automated. Next step is to analyze and create visualizations with the transformed dataset using Power BI tool. I will create connection in Power BI with BigQuery dataset and establish the live connection. this way visuals will be automatically refreshed when the Power BI refresh is done. Overall goal is to implement an automated pipeline with maximizing the use of tools and concepts learnt in INFO-I535 course.

BACKGROUND:

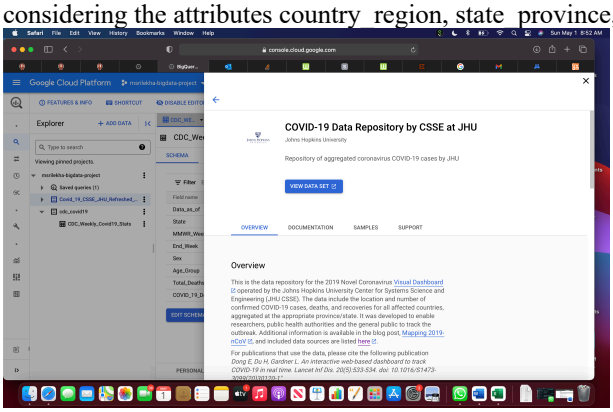
The problem statement was selected to implement the pipeline using Google Cloud Platform. This problem statement addresses use of multiple concepts present in the course modules, like Data types, data world, Data pipeline and lifecycle, GCP, Qwiklabs sessions related to Google cloud, data visualization, Ingestion and storage, Data modelling. I have selected this project because GCP is great platform to ingest data, store in cloud, transform using BigQuery. Any complex dataset can be worked on in GCP, with automated query refreshes and establish connections with external visualization tools. This project outcome will be an ETL automated pipeline with learning outcomes like working with GCP, BigQuery and Power BI.
The dataset I have selected for this project is COVID – 19 cases in the world which is a public dataset created by Johns Hopkins University. This data is updated every day and is at date level. I want to aggregate the data to month level as viewing the data and reports at day level will be very complex and not insightful. I want to create insights of the datapoints like Cases confirmed and deaths across the country, state and month and year level. Converting the dataset from day level to month level makes the problem statement interesting and important, and using GCP and BigQuery to obtain the results.

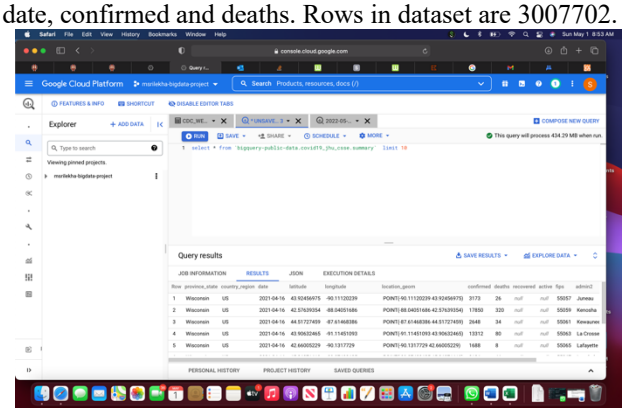
METHOLODOGY:

I started with setting up the environment. I have redeemed the coupon code and created new project named *msrilekha-bigdata-project* in Google cloud platform.

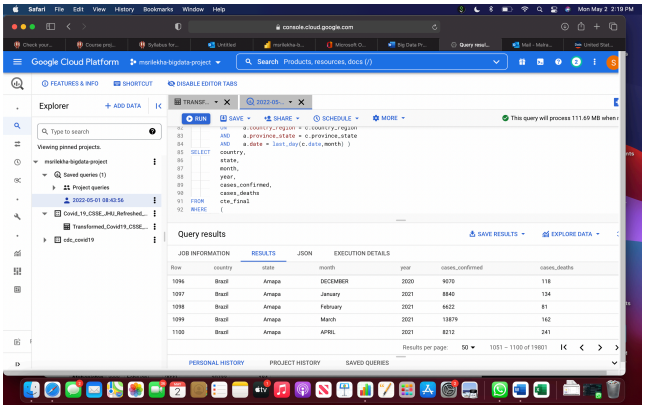


Next step is to create the dataset, I have selected the public dataset in GCP from marketplace. Covid-19 Data repository by CSSE at JHU. I have viewed the dataset and schema of the dataset. Below is the sample of the data present. Brief explanation of the dataset and attributed is as follows:
This dataset contains location and number of confirmed COVID-19 cases and deaths for affected countries aggregated at the appropriate state/province level. This dataset refreshes daily and data for every day since the covid-19. I will be considering the attributes country_region, state_province, date, confirmed and deaths. Rows in dataset are 3007702.

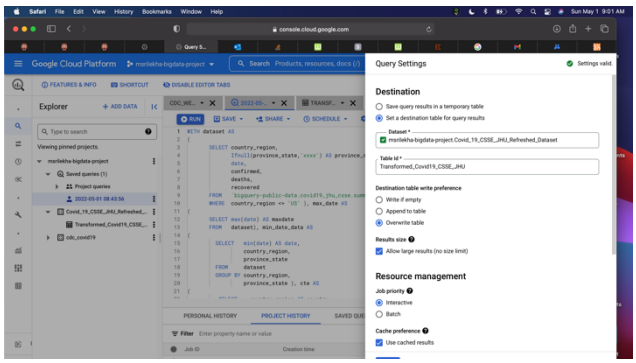
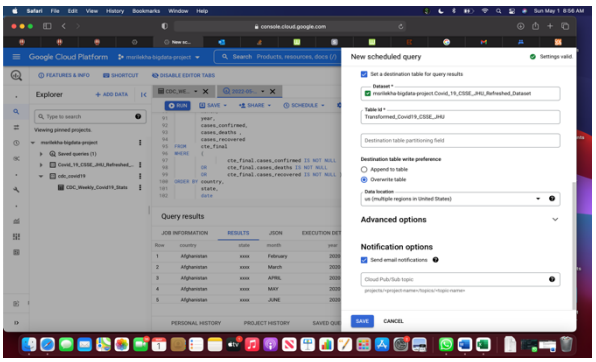
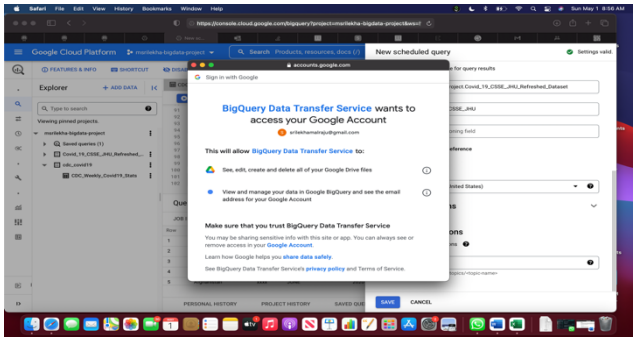
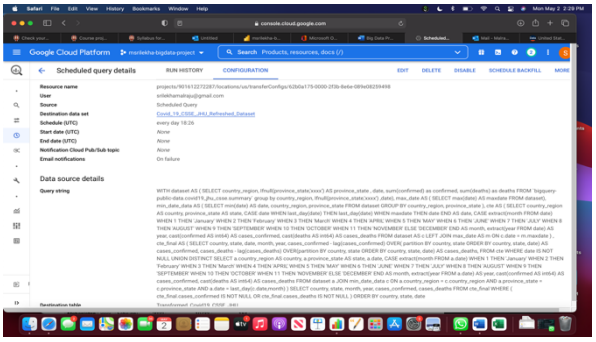




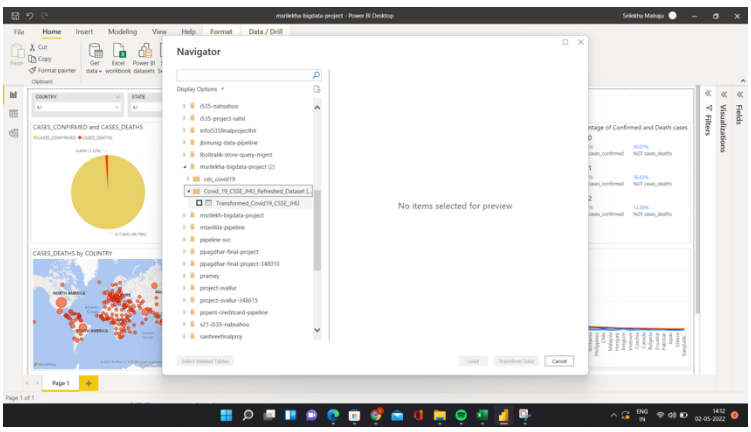
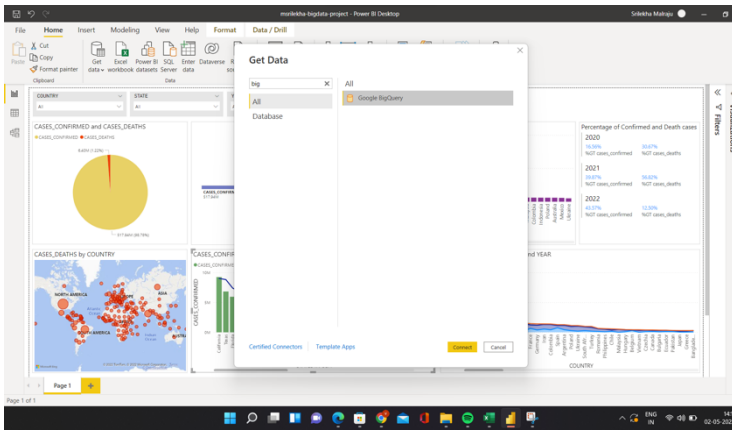
Now that dataset is ready, I have proceeded with Transforming the data and prepare a table with required data and attributes. BigQuery is used in this step. Firstly, I have extracted the year and month from the date column available. As the data at date level is cumulative which means if suppose as of yesterday cases are 500, and cases today are 50, then confirmed cases will be 550 for today's date, similarly for deaths. Hence to calculate cases data for a particular month I have calculated last day of each month for each country, state and year, and hence august month cases will be for year 2021 = (August 31, 2021 - July 31, 2021) cases and aggregated at country, state, and year level. If there is no previous month, then month cases will be the data on last day of the month. As data for USA is at different county level, I have aggregated data to be at state and country level.
Also, I have eliminated the rows which don't have any confirmed and deaths.



Next, I have created dataset table to store the query results. Further, I have created schedule query. I have enabled the API and schedule the query for dataset and table. This will automate the process of refreshing the dataset. Dumped the transformed data to the table.

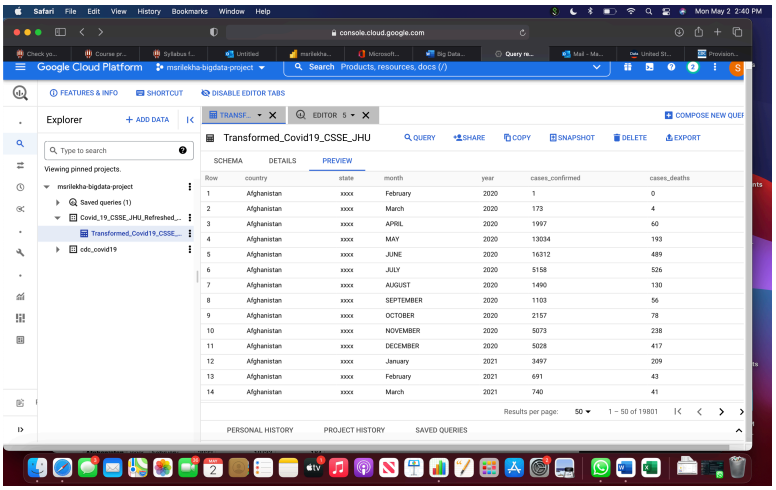





Connecting the BigQuery with PowerBI to visualize the transformed data. Developed appropriate charts like Confirmed and Deaths and cases (Numbers and percentage), Decomposition tree of confirmed cases across years and months. Confirmed cases bar visual by country, deaths reported in different countries map visual.

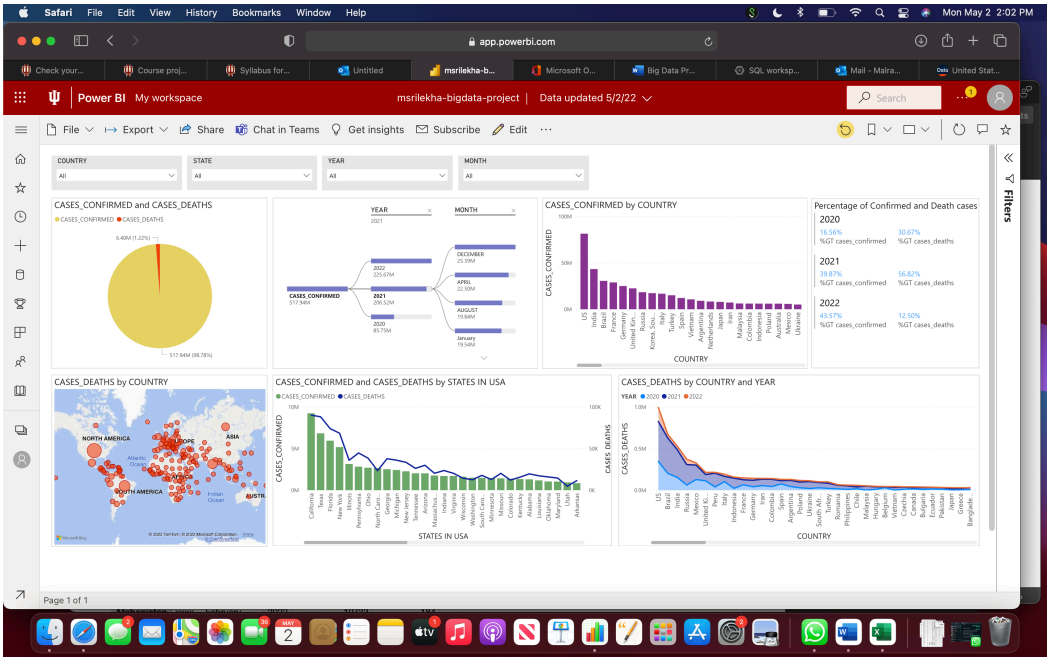



RESULTS:

Results of the methodology, execution of query resulted in transformed data as desired. The transformed data has month wise data for all countries and states. I have eliminated unused attributes, rows which have null values in all cases confirmed and deaths. Transformed data has only 19801 rows because the aggregation is not at date level and is at month level. Below is the preview of the results table.



I have visualized the data in the Power BI as it is insightful and informative to view data in power bi visuals. I have added filters to the report so that any user can select required month or year or state or country for which they want to view the COVID 19 cases. Below are the different kind of visuals I have created as per data requirements.



We can see that only 1.22% of deaths has occurred whereas confirmed cases are 98.76% from the first chart. In the second visual we can get insights on confirmed cases for each month and each year across world. In third visual we can see that confirmed cases are more in USA, India, Brazil. Next chart depicts the percentage of total cases and deaths in for 2020,2021 and 2022. We can see that 2021 has most deaths and death rate is highly reduced when it comes to 2022. In the map chart we can see that large clusters have more deaths where USA, India has highest deaths, next we are seeing cases and deaths in different states of USA where highly effected states are California, Texas, Florida. The last visual shows deaths by country and year.

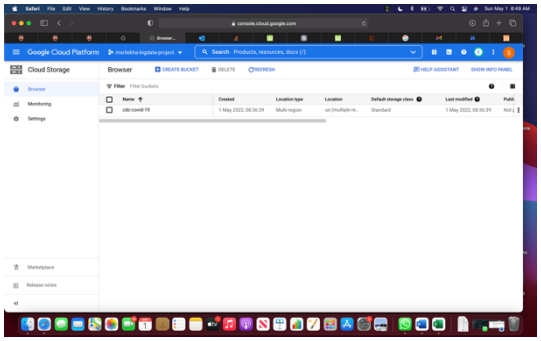
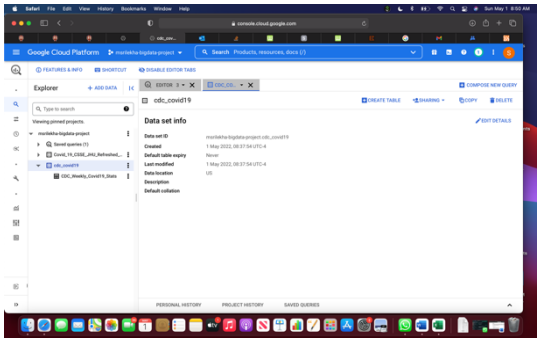
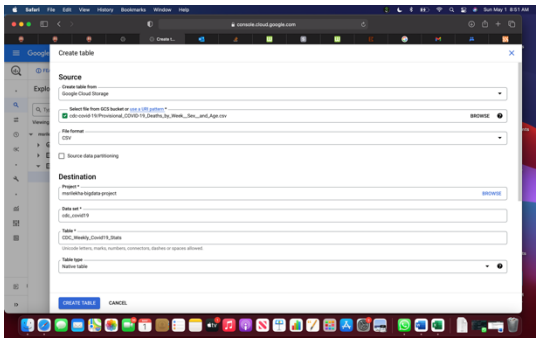
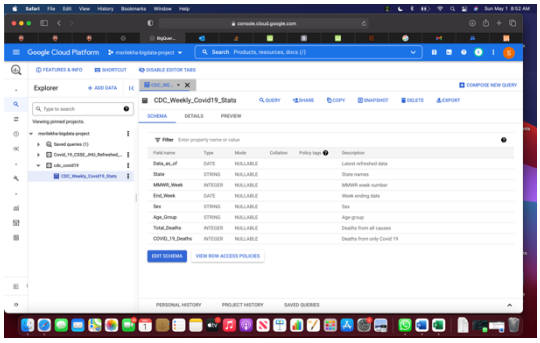
Published report link:

https://app.powerbi.com/view?r=eyJrIjoiYWYyOTlhMTUtYWVkZC00ZDg0LTliY2UtZWJhMTMzYTU0NDY2IiwidCI6IjExMTNiZTM0LWFlZDEtNGQwMC1hYjRiLWNkZDAyNTEwYmU5MSIsImMiOjN9

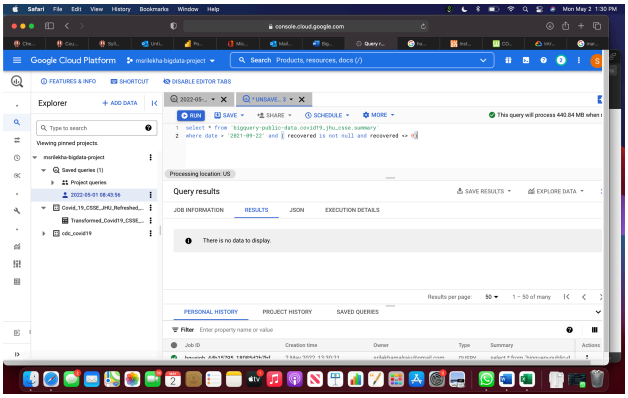
Hence, from this report we can get insights about the COVID 19 cases in different countries, states, month, and year. I have also used all the technologies and tools which were described the introduction and background and built an automated ETL pipeline using GCP and BigQuery.

DISCUSSION:

We can interpret from the visuals clearly number of confirmed cases and deaths around the world and in different countries and states. We can also see the comparison of all cases across the countries, states. WE can also find out where the covid has major impact, in which year month, country or state.
I have employed the technologies and tools from this course in building the ETL pipeline. I have used the GCP platform using the coupon code, used public dataset from the GCP marketplace, performed transformations in BigQuery, created schedule queries, and connected the BigQuery dataset and table with external visualization tool PowerBI.
Barriers or failures included: Initially I thought of choosing dataset from CDC and implemented it by manually (https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-Week-Sex-and-Age/vsak-wrfu) downloading the dataset and created a cloud storage bucket and uploaded the csv file. And then created dataset and added table in BigQuery to perform further transformation. But then I realized in this case the dataset cannot be automated as the data is in my local machine, hence I changed my approach and selected public dataset.

Another barrier is Confirmed, and death cases are present till today and consistent but recovered cases data is present only till August 2021 and inconsistent for all countries in the dataset, hence I am not showing recovered cases in this project. This project shows only confirmed and deaths cases stats.



CONCLUSION:

The problem statement was to implement the ETL automated pipeline using Google Cloud Platform. I have implemented ETL pipeline using GCP, ingested and extracted data, performed transformation using BigQuery, automated the dataset refresh procedure, connected BigQuery with external visualization tool and visualized the results. From the PowerBI dashboard I have developed we can get insights about COVID 19 cases for different countries, states for each month and year. This project was a great learning experience, and I got to explore and learn the concepts which were beyond the course work. Working with cloud platforms is always a great experience as its very interactive and has minimal resources or architecture bounds.

REFERENCES:

https://data.cdc.gov/
https://cloud.google.com/bigquery/public-data/
https://cloud.google.com/bigquery/docs/reference/standard-sql/conversion_functions
https://cloud.google.com/learn/what-is-etl
https://cloud.google.com/bigquery/docs/scheduling-queries
https://docs.microsoft.com/en-us/power-bi/connect-data/desktop-connect-bigquery
and CANVAS Materials