Implementation Project

on

COVID-19
BIG-DATA ANALYSIS

Course: CPSC 531- Advanced Database Management
Section: 3
Team Members:

Shriya Bannikop	**885196238**
Debdyuti Das	**886676550**

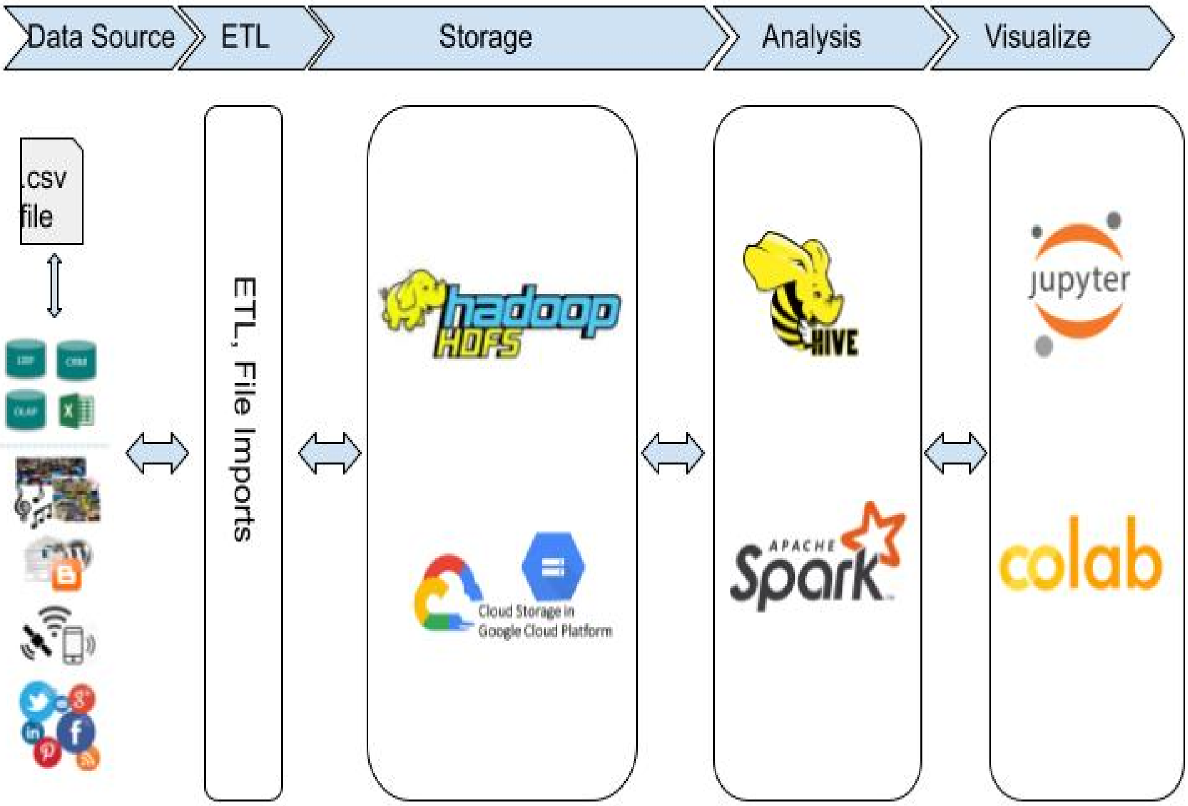
PROBLEM STATEMENT

The risk of coronavirus was still increasing even after the government had taken several measures worldwide to minimise the Covid-19 spread. The transmission chain's severity was deemed broken only when no new case was reported in an area. The only way to break the transmission chain is to impose a Lockdown.

The aim of the project is to address, compare and analyse the variation in the number of COVID-19 cases in countries which imposed complete lockdown with restriction rules and observe the following trend:

- To capture the trend in the data based on the increasing number of cases
- Was imposing lockdown a right decision
- Compare countries which imposed lockdown and analyse the variation in the number of covid 19 cases

ARCHITECTURE



- Big Data Technologies utilised in Cluster created on Google Cloud Platform i.e. Dataproc
- .csv files are given as input
- Data ingestion is done into Hadoop Distributed File System (HDFS) and stored into Google Cloud Storage Bucket
- Extract Data in Hive,Spark for analysis
- Extracted data using Hive (Hive is used as ETL to connect HDFS and spark) and used Apache Spark to perform the analysis
- The output of the analysed data is visualised using Jupyter Notebook
- The files are stored back into google storage bucket

TOOLS AND TECHNOLOGIES

- **Cloud Platform**: Google Cloud Cluster
- **Primary Storage System**: Hadoop Distributed File System
- **Distributed processing System**: Apache spark
- **ETL Tools**: Apache Hive
- **Visualisation**: Jupyter Notebook, Google Colab

FUNCTIONALITIES

- To capture the trend in the data collected from multiple datasets based on the increasing number of cases
- To determine if imposing lockdown was a right decision
- To compare countries which imposed lockdown and analyse the variation in the number of Covid-19 cases
- Migration analysis to know the population and cases before and after lockdown

APPROACH

Before implementing with Cluster:

1. Download files and store in local directory
2. Start all daemons in HDFS using
   - `hdfs namenode -format`
   - `start-dfs.sh`
3. Verify if all components are running
   - `jps`
4. Move .csv files to HDFS
   - Make a directory in HDFS:
     ```
     hadoop fs -mkdir -p /home/hadoop/directory_name
     ```
   - Copy the .csv file from Local to HDFS:
     ```
     hadoop fs -put
     /home/debdyuti/bigdata/covid_19_data.csv
     /home/hadoop/directory_name
     ```
   - Check if its copied:
     ```
     hadoop fs -ls /home/hadoop/directory_name
     ```
5. Create tables in hive and use MapReduce
   - `Cd $HIVE_HOME/bin`
   - Open hive-CLI: `hive`
   - Create database:
     ```
     CREATE SCHEMA IF NOT EXISTS database_name;
     USE database_name;
     ```
   - Create table:
     ```
     CREATE TABLE IF NOT EXISTS
     database_name.covid_details(SNo INTEGER,
     ObservationDate STRING, State STRING, Country
     STRING, LastUpdate STRING, Confirmed DOUBLE,
     Deaths DOUBLE, Recovered DOUBLE)
     ROW FORMAT DELIMITED
     FIELDS TERMINATED BY ',';
     ```

- Load Dataset from HDFS to HIVE Table:
  ```
  LOAD DATA INPATH
  '/home/hadoop/directory_name/covid_19_data.csv'
  INTO TABLE database_name.covid_details ;
  ```
- To see records in the HIVE table:
  ```
  SELECT * FROM database_name.covid_details ;
  ```

6. Extract data into Spark for analysis
7. Read files from Spark and visualise using Google Colab

After implementing with Cluster:

1. Create cluster in Google Cloud Platform
2. Open console (SSH) on master node
3. Download data from internet into the Hadoop cluster (HDFS location) using `wget` command
4. Copy files from hadoop (HDFS location) into Google storage bucket using "`gsutil cp migration_population.csv us-central1 gs://dyutishriya-bucketdbms/Data1;`"
5. Using web interface analyse and data in Jupyter Notebook
6. Read from Google Storage Bucket. After analysis and visualisation store it back into the bucket.

STEPS TO RUN THE PROJECT

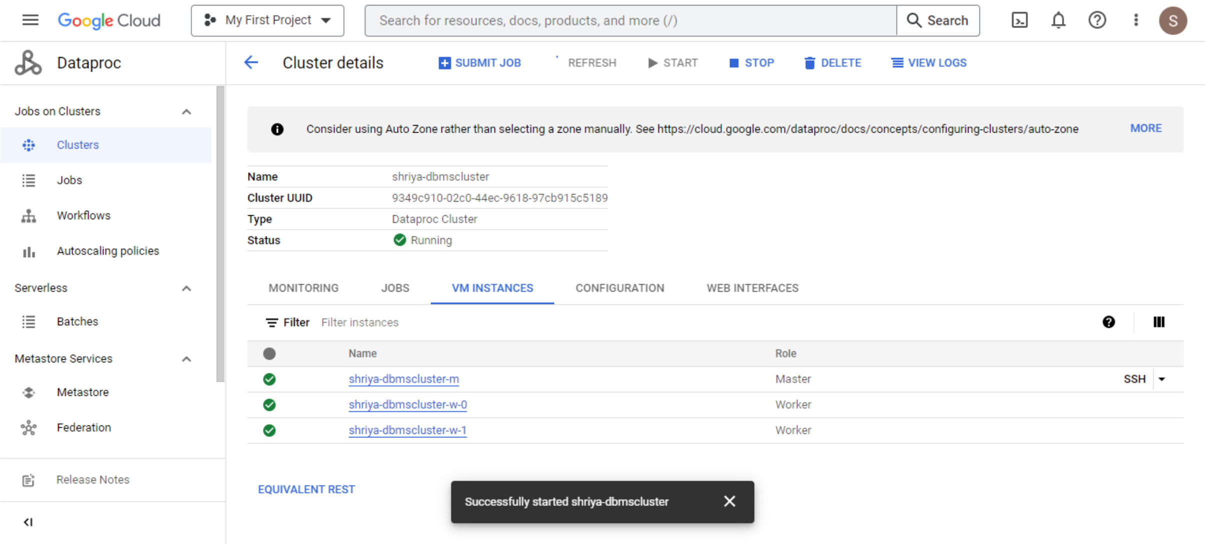
Github Location of Code :

https://github.com/Debdyuti-01/Covid-19-Big-Data-Analysis

1. Start the Cluster
2. Start demons (start-dfs command) by opening SSH shell in master node
3. Run the .ipynb file on jupyter notebook to check visualisations

TEST RESULTS OF SPARK ANALYSIS

- ## Cluster Creation:



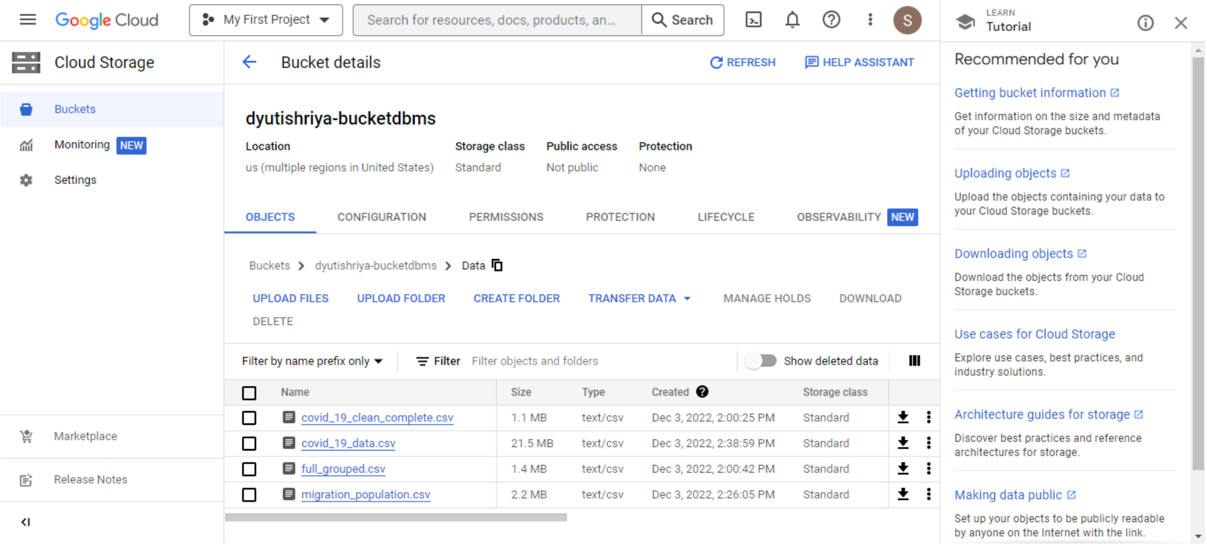
Consider using Auto Zone rather than selecting a zone manually. See https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/auto-zone

Name	shriya-dbmscluster
Cluster UUID	9349c910-02c0-44ec-9618-97cb915c5189
Type	Dataproc Cluster
Status	✓ Running

	Name	Role
✓	shriya-dbmscluster-m	Master
✓	shriya-dbmscluster-w-0	Worker
✓	shriya-dbmscluster-w-1	Worker

Successfully started shriya-dbmscluster

- ## Bucket Creation and Loading data:



dyutishriya-bucketdbms

Location	Storage class	Public access	Protection
us (multiple regions in United States)	Standard	Not public	None

Buckets > dyutishriya-bucketdbms > Data

	Name	Size	Type	Created	Storage class
□	covid_19_clean_complete.csv	1.1 MB	text/csv	Dec 3, 2022, 2:00:25 PM	Standard
□	covid_19_data.csv	21.5 MB	text/csv	Dec 3, 2022, 2:38:59 PM	Standard
□	full_grouped.csv	1.4 MB	text/csv	Dec 3, 2022, 2:00:42 PM	Standard
□	migration_population.csv	2.2 MB	text/csv	Dec 3, 2022, 2:26:05 PM	Standard

Recommended for you

Getting bucket information
Get information on the size and metadata of your Cloud Storage buckets.

Uploading objects
Upload the objects containing your data to your Cloud Storage buckets.

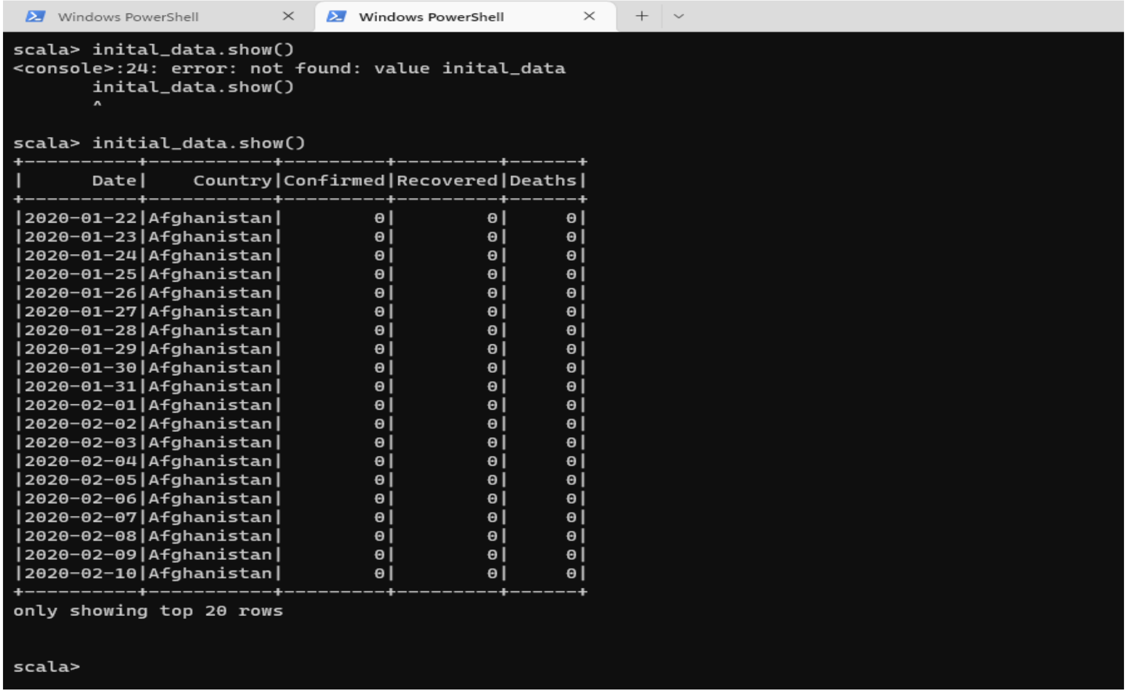
Downloading objects
Download the objects from your Cloud Storage buckets.

Use cases for Cloud Storage
Explore use cases, best practices, and industry solutions.

Architecture guides for storage
Discover best practices and reference architectures for storage.

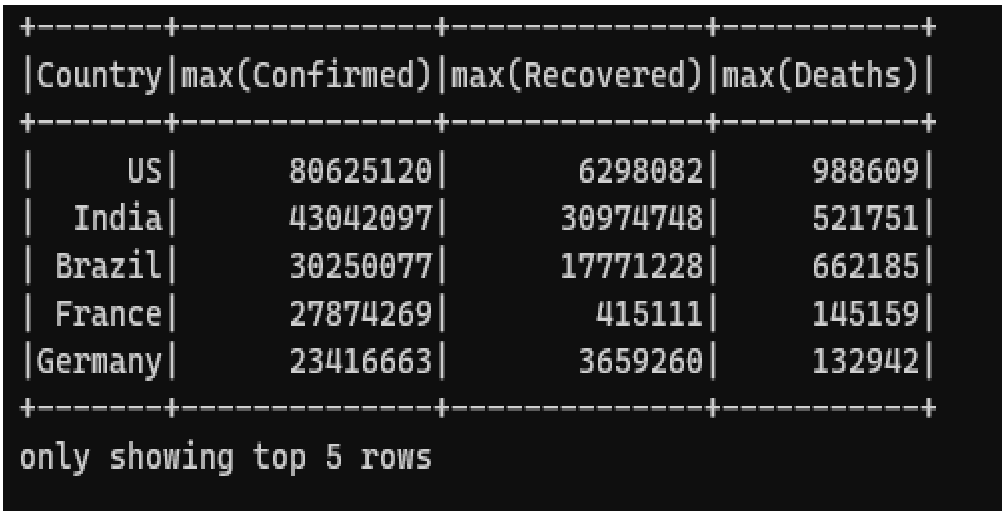
Making data public
Set up your objects to be publicly readable by anyone on the Internet with the link.

- **Data Cleaning:**
    - Filled blank fields with ‘unknown’
    - Filtered data
    - Converted String datatype to Date datatype of Date attribute

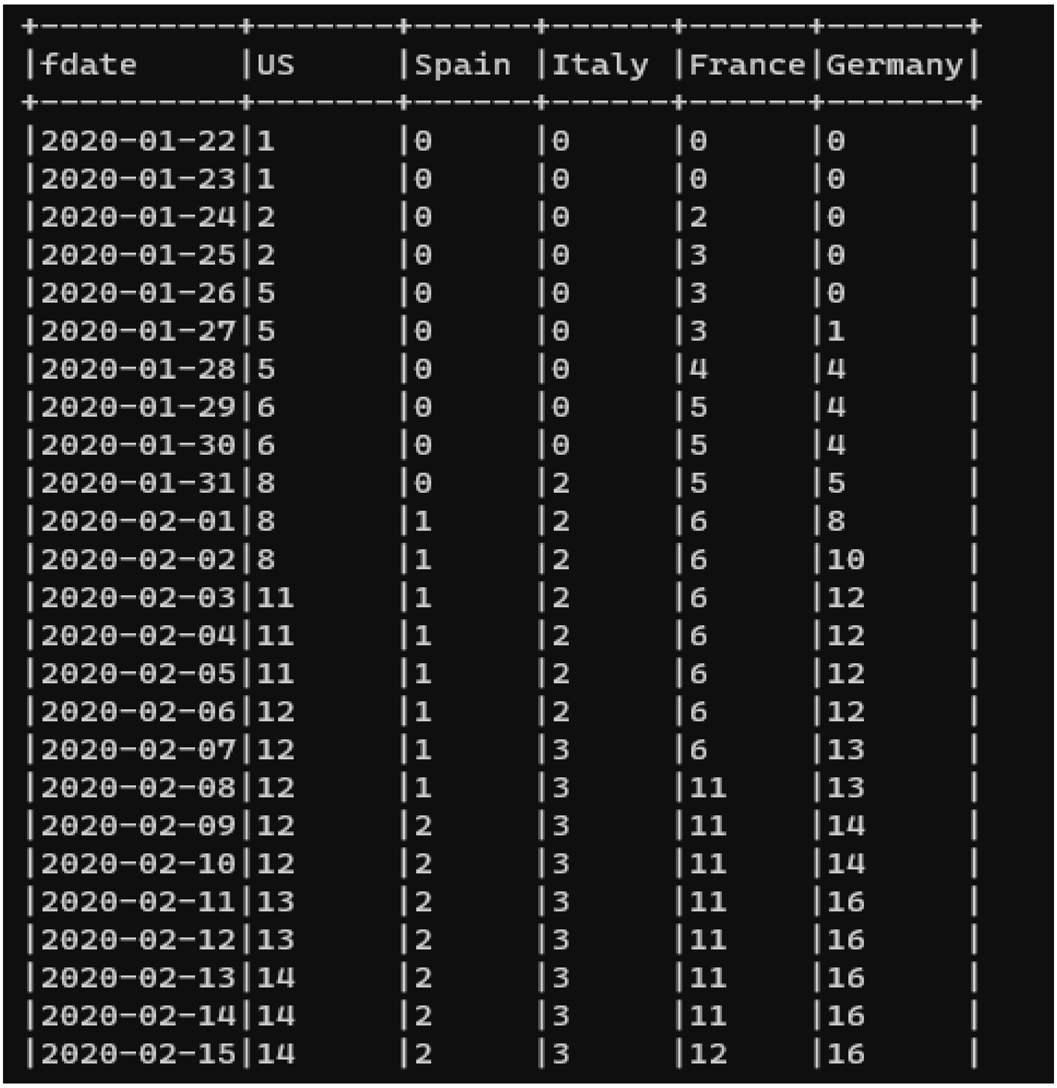


```
scala> inital_data.show()
<console>:24: error: not found: value inital_data
       inital_data.show()
       ^

scala> initial_data.show()
+----------+-----------+---------+---------+------+
|      Date|    Country|Confirmed|Recovered|Deaths|
+----------+-----------+---------+---------+------+
|2020-01-22|Afghanistan|        0|        0|     0|
|2020-01-23|Afghanistan|        0|        0|     0|
|2020-01-24|Afghanistan|        0|        0|     0|
|2020-01-25|Afghanistan|        0|        0|     0|
|2020-01-26|Afghanistan|        0|        0|     0|
|2020-01-27|Afghanistan|        0|        0|     0|
|2020-01-28|Afghanistan|        0|        0|     0|
|2020-01-29|Afghanistan|        0|        0|     0|
|2020-01-30|Afghanistan|        0|        0|     0|
|2020-01-31|Afghanistan|        0|        0|     0|
|2020-02-01|Afghanistan|        0|        0|     0|
|2020-02-02|Afghanistan|        0|        0|     0|
|2020-02-03|Afghanistan|        0|        0|     0|
|2020-02-04|Afghanistan|        0|        0|     0|
|2020-02-05|Afghanistan|        0|        0|     0|
|2020-02-06|Afghanistan|        0|        0|     0|
|2020-02-07|Afghanistan|        0|        0|     0|
|2020-02-08|Afghanistan|        0|        0|     0|
|2020-02-09|Afghanistan|        0|        0|     0|
|2020-02-10|Afghanistan|        0|        0|     0|
+----------+-----------+---------+---------+------+
only showing top 20 rows


scala>
```

- **Data Exploration:**
    - To find top 5 countries which were leading with Covid-19 cases
    - Pivoted the table by Country attribute
    - Total number of recovered, confirmed and death cases of Covid of top 5 leading countries

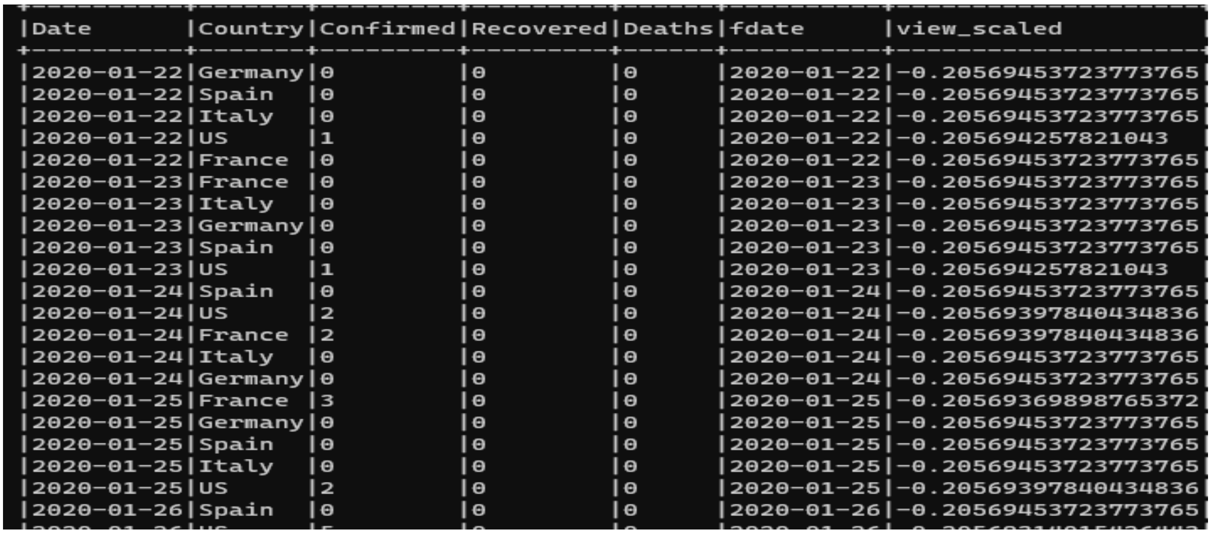


```
+-------+--------------+--------------+-----------+
|Country|max(Confirmed)|max(Recovered)|max(Deaths)|
+-------+--------------+--------------+-----------+
|     US|      80625120|       6298082|     988609|
|  India|      43042097|      30974748|     521751|
| Brazil|      30250077|      17771228|     662185|
| France|      27874269|        415111|     145159|
|Germany|      23416663|       3659260|     132942|
+-------+--------------+--------------+-----------+
only showing top 5 rows
```



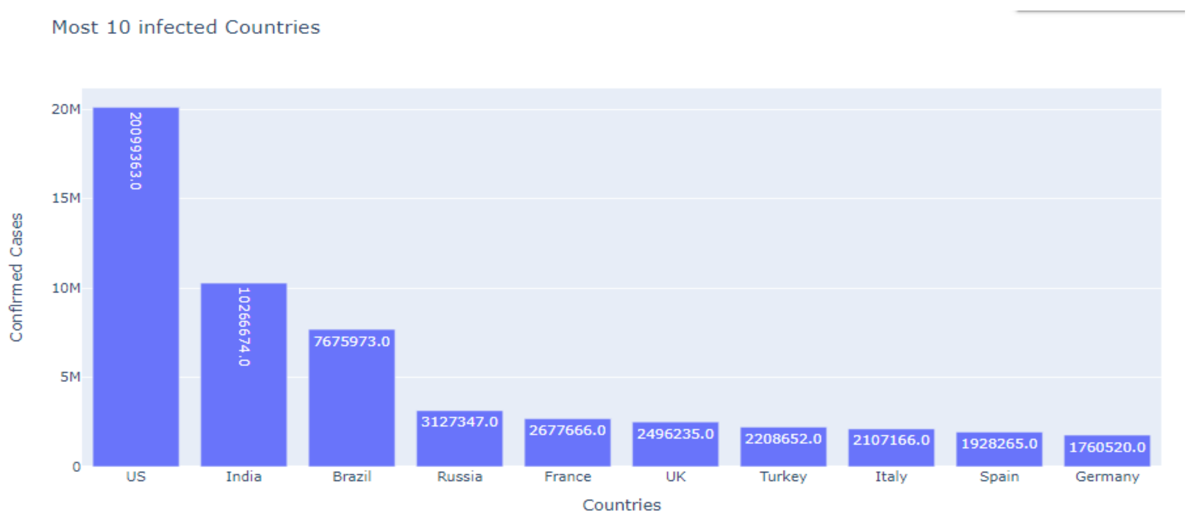
```
+----------+-------+------+------+------+-------+
|fdate     |US     |Spain |Italy |France|Germany|
+----------+-------+------+------+------+-------+
|2020-01-22|1      |0     |0     |0     |0      |
|2020-01-23|1      |0     |0     |0     |0      |
|2020-01-24|2      |0     |0     |2     |0      |
|2020-01-25|2      |0     |0     |3     |0      |
|2020-01-26|5      |0     |0     |3     |0      |
|2020-01-27|5      |0     |0     |3     |1      |
|2020-01-28|5      |0     |0     |4     |4      |
|2020-01-29|6      |0     |0     |5     |4      |
|2020-01-30|6      |0     |0     |5     |4      |
|2020-01-31|8      |0     |2     |5     |5      |
|2020-02-01|8      |1     |2     |6     |8      |
|2020-02-02|8      |1     |2     |6     |10     |
|2020-02-03|11     |1     |2     |6     |12     |
|2020-02-04|11     |1     |2     |6     |12     |
|2020-02-05|11     |1     |2     |6     |12     |
|2020-02-06|12     |1     |2     |6     |12     |
|2020-02-07|12     |1     |3     |6     |13     |
|2020-02-08|12     |1     |3     |11    |13     |
|2020-02-09|12     |2     |3     |11    |14     |
|2020-02-10|12     |2     |3     |11    |14     |
|2020-02-11|13     |2     |3     |11    |16     |
|2020-02-12|13     |2     |3     |11    |16     |
|2020-02-13|14     |2     |3     |11    |16     |
|2020-02-14|14     |2     |3     |11    |16     |
|2020-02-15|14     |2     |3     |12    |16     |
```

- **Data Preparation**
    - Scaled values to remove outliers

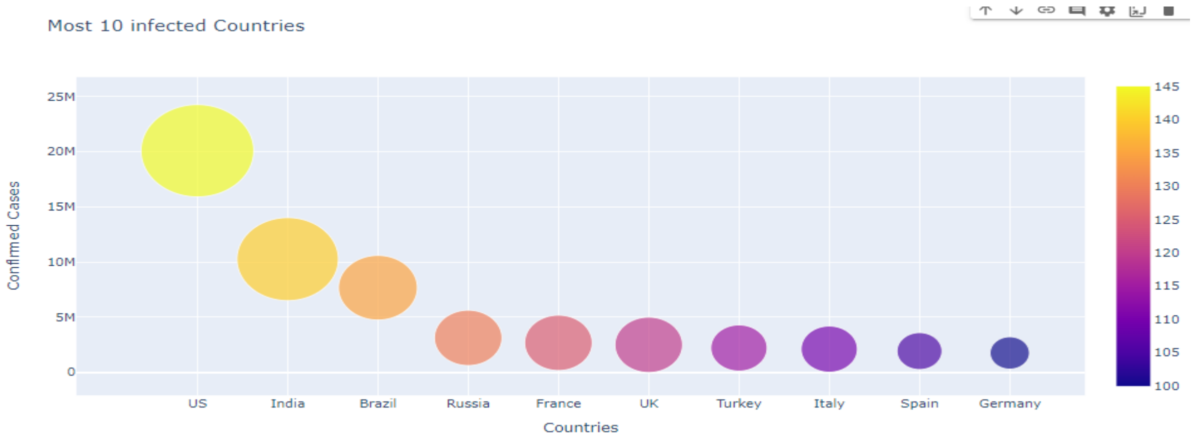


```
+----------+-------+---------+---------+------+----------+--------------------+
|Date      |Country|Confirmed|Recovered|Deaths|fdate     |view_scaled         |
+----------+-------+---------+---------+------+----------+--------------------+
|2020-01-22|Germany|0        |0        |0     |2020-01-22|-0.20569453723773765|
|2020-01-22|Spain  |0        |0        |0     |2020-01-22|-0.20569453723773765|
|2020-01-22|Italy  |0        |0        |0     |2020-01-22|-0.20569453723773765|
|2020-01-22|US     |1        |0        |0     |2020-01-22|-0.205694257821043  |
|2020-01-22|France |0        |0        |0     |2020-01-22|-0.20569453723773765|
|2020-01-23|France |0        |0        |0     |2020-01-23|-0.20569453723773765|
|2020-01-23|Italy  |0        |0        |0     |2020-01-23|-0.20569453723773765|
|2020-01-23|Germany|0        |0        |0     |2020-01-23|-0.20569453723773765|
|2020-01-23|Spain  |0        |0        |0     |2020-01-23|-0.20569453723773765|
|2020-01-23|US     |1        |0        |0     |2020-01-23|-0.205694257821043  |
|2020-01-24|Spain  |0        |0        |0     |2020-01-24|-0.20569453723773765|
|2020-01-24|US     |2        |0        |0     |2020-01-24|-0.20569397840434836|
|2020-01-24|France |2        |0        |0     |2020-01-24|-0.20569397840434836|
|2020-01-24|Italy  |0        |0        |0     |2020-01-24|-0.20569453723773765|
|2020-01-24|Germany|0        |0        |0     |2020-01-24|-0.20569453723773765|
|2020-01-25|France |3        |0        |0     |2020-01-25|-0.20569369898765372|
|2020-01-25|Germany|0        |0        |0     |2020-01-25|-0.20569453723773765|
|2020-01-25|Spain  |0        |0        |0     |2020-01-25|-0.20569453723773765|
|2020-01-25|Italy  |0        |0        |0     |2020-01-25|-0.20569453723773765|
|2020-01-25|US     |2        |0        |0     |2020-01-25|-0.20569397840434836|
|2020-01-26|Spain  |0        |0        |0     |2020-01-26|-0.20569453723773765|
```

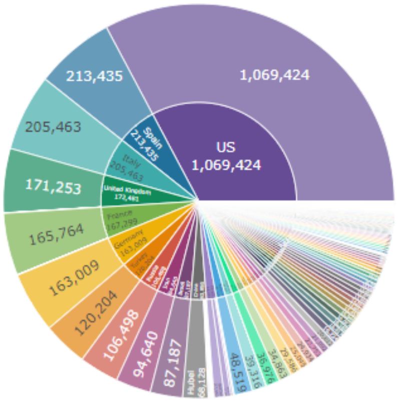
- **Analysis**



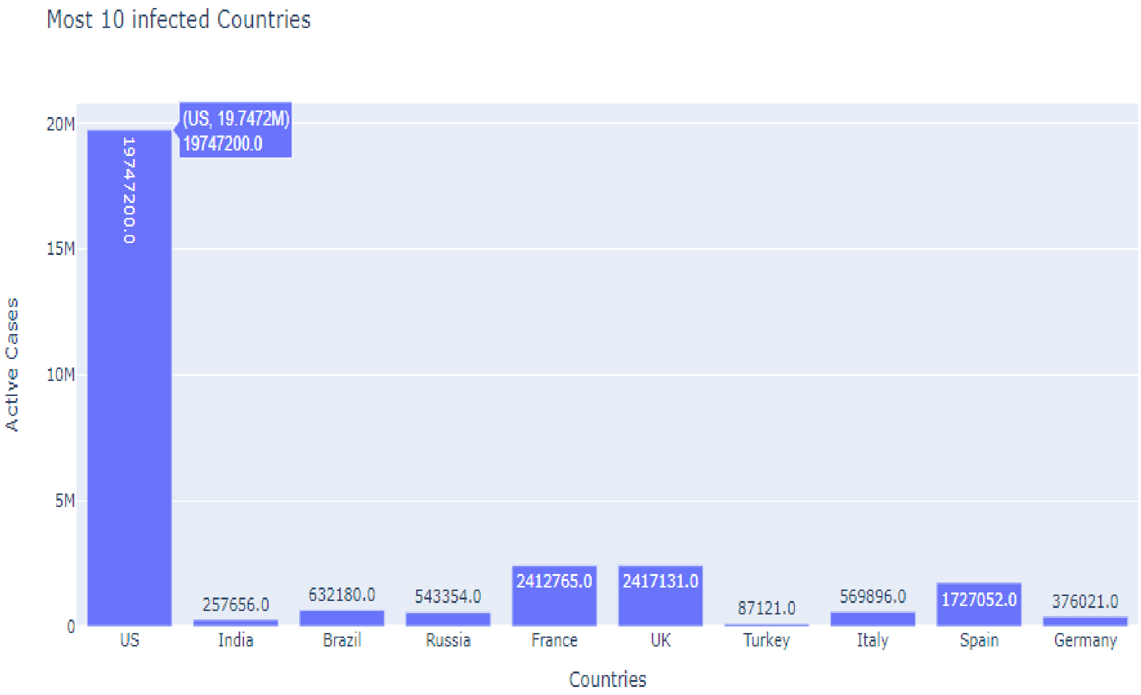
Confirmed covid cases in year 2020 for each country

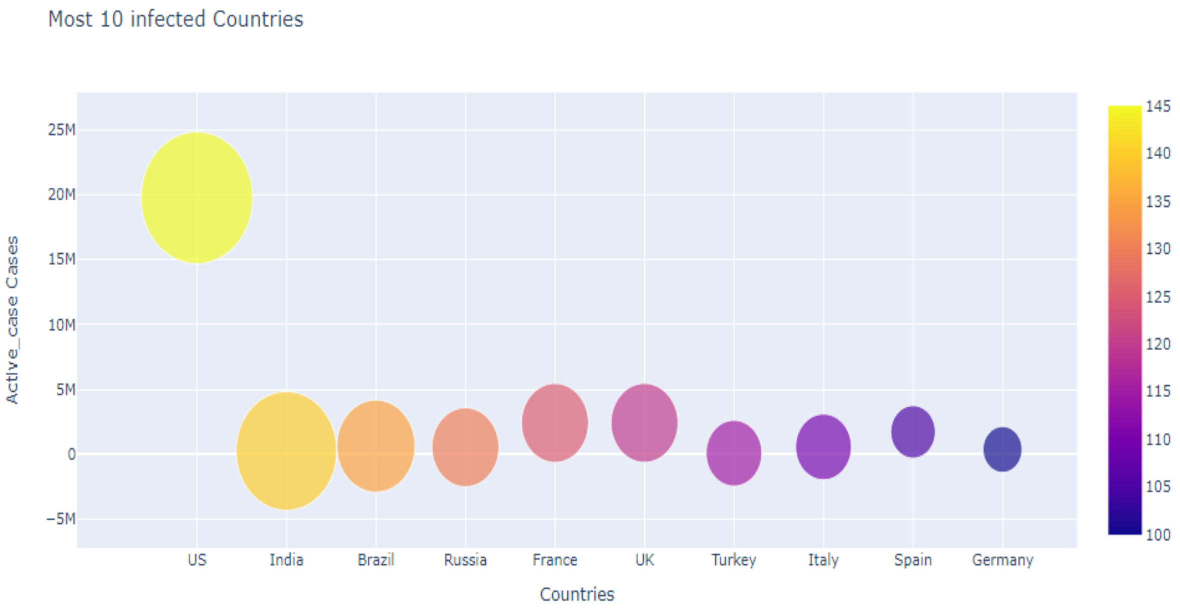


Top 10 countries sorted by maximum number of confirmed covid cases

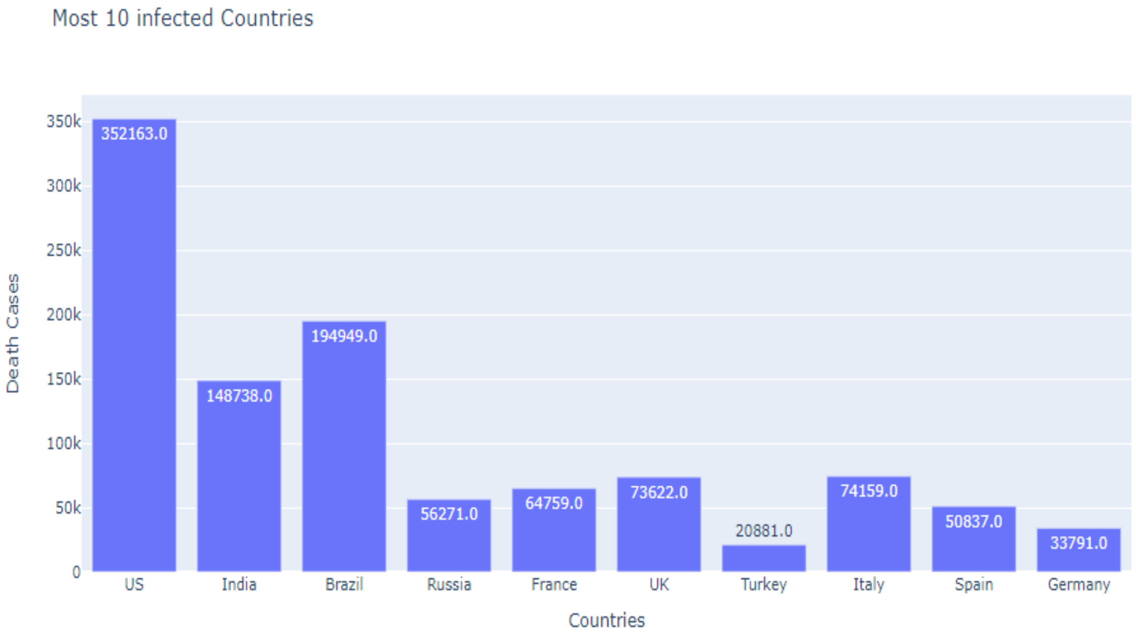


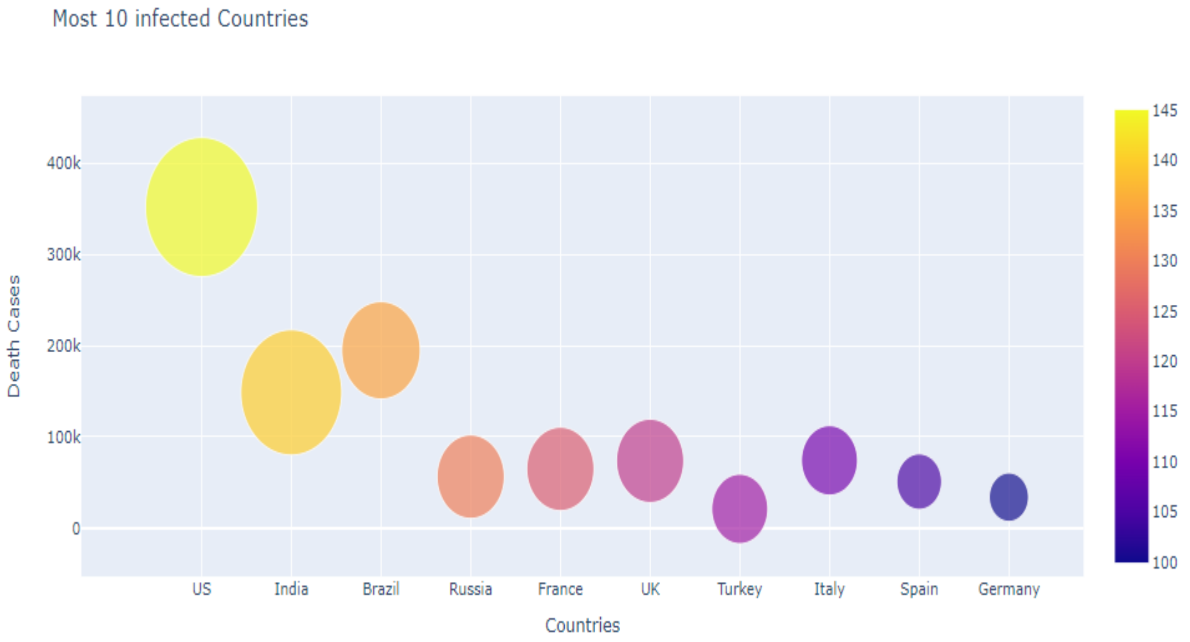
- Total number of active cases in each country
- Top 10 countries sorted by number of active covid cases



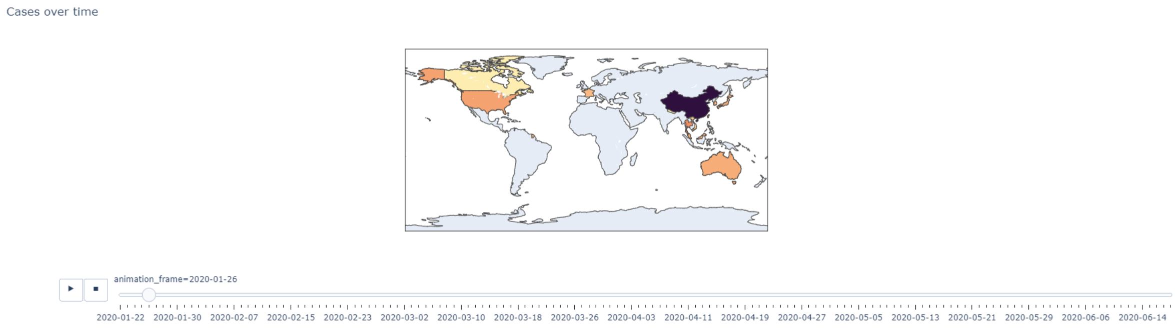


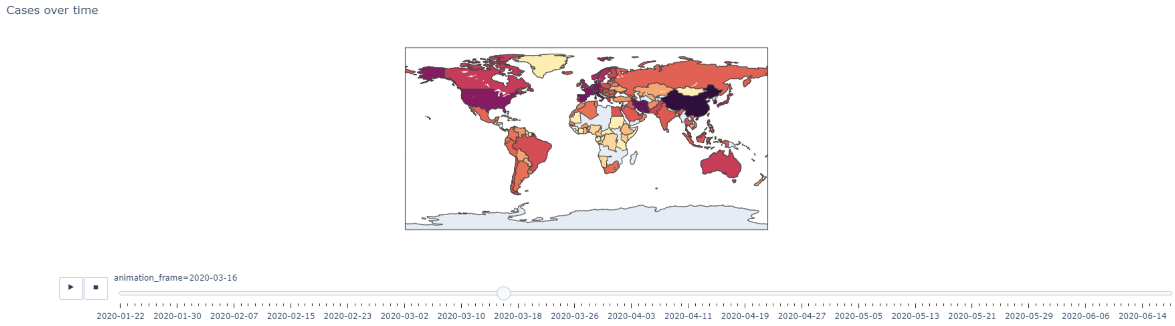
- Total number of death cases in each country
- Top 10 country sorted by covid death rate

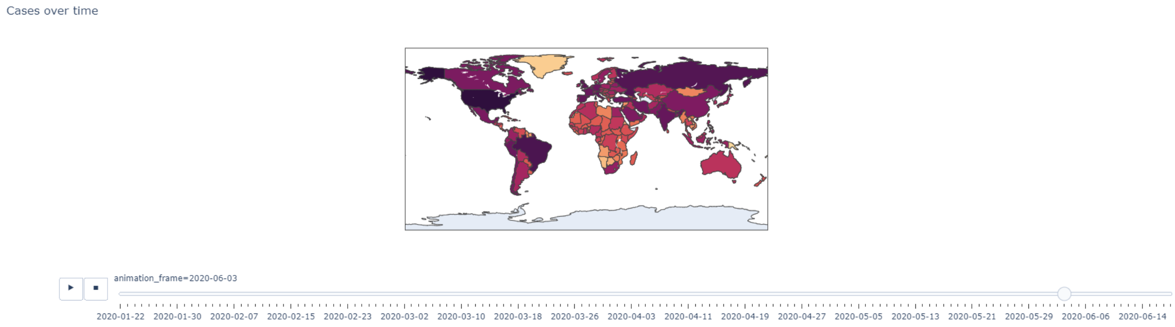




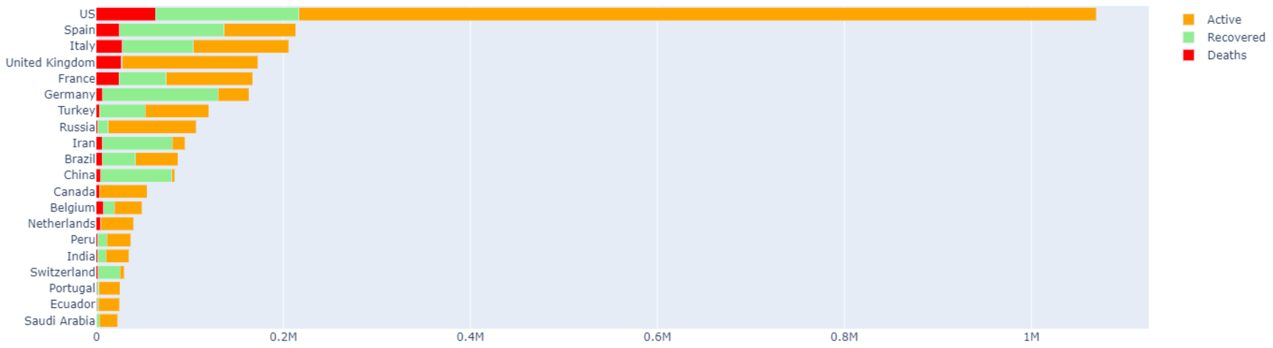
- Plot showing the increases in cases by date(animation plot)



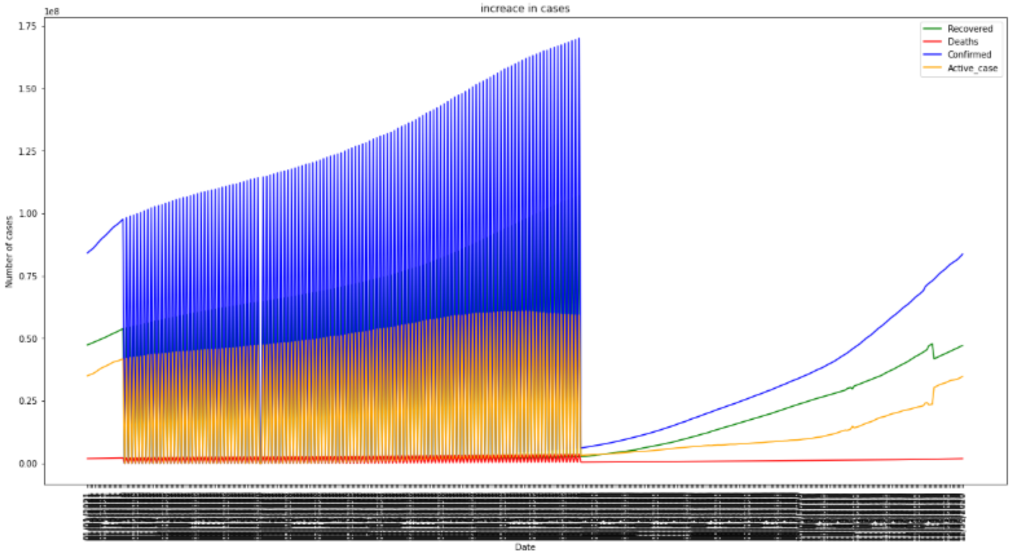




- Top 20 countries sorted based on the total number of cases

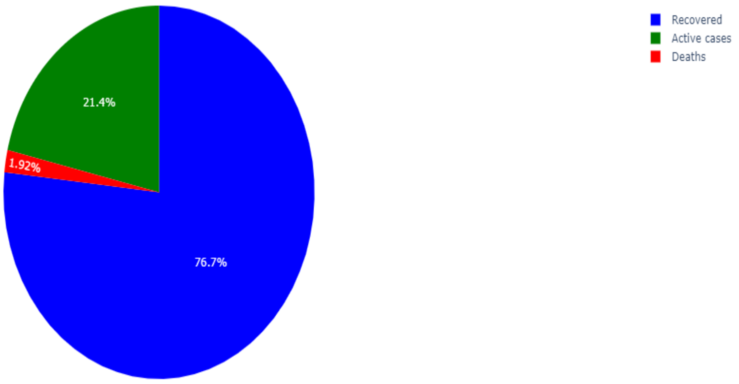


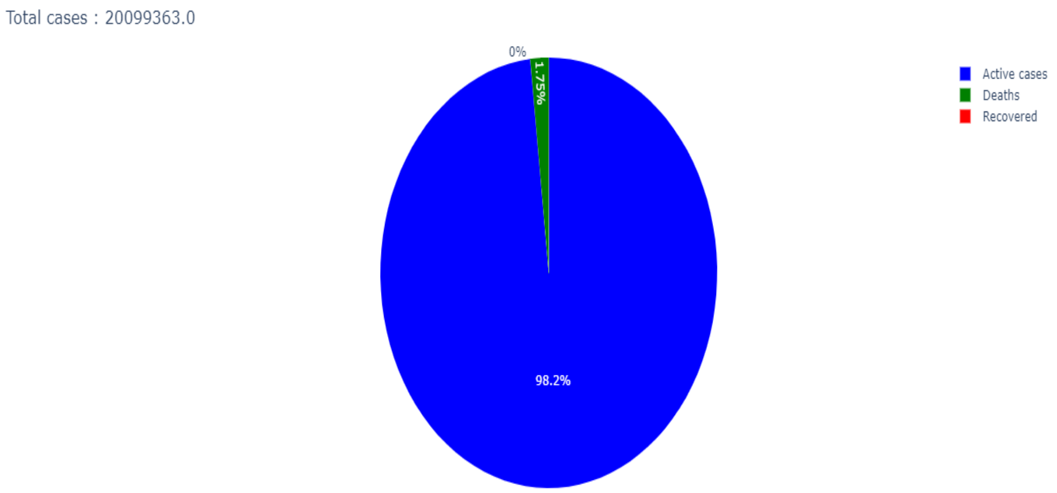
- Line Chart of increase in 'Recovered', 'Deaths', 'Confirmed', 'Active_case'

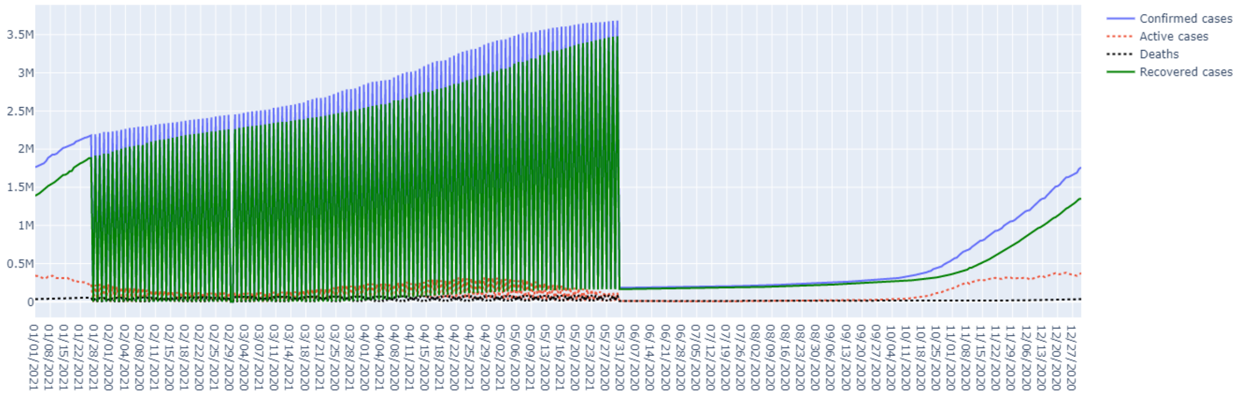


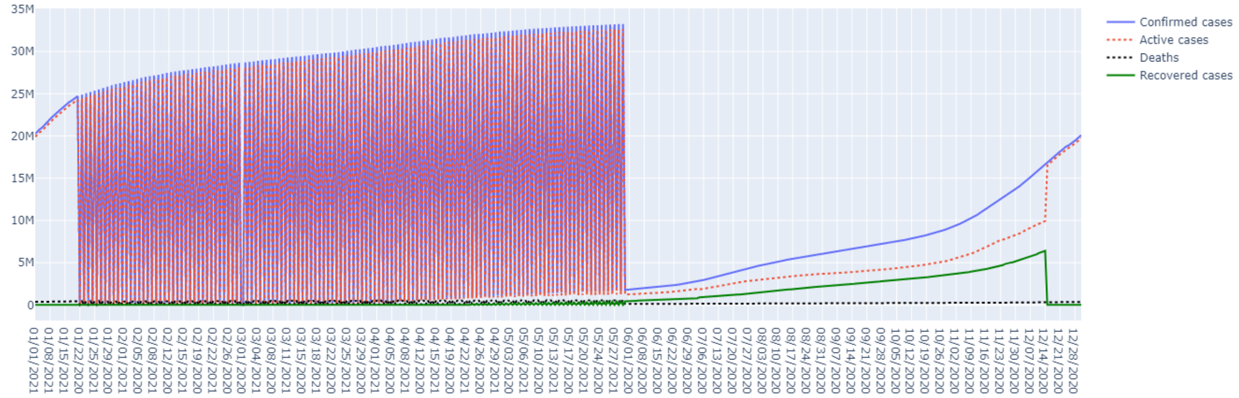
- **Comparisons**
  Germany vs USA

Total cases : 1760520.0





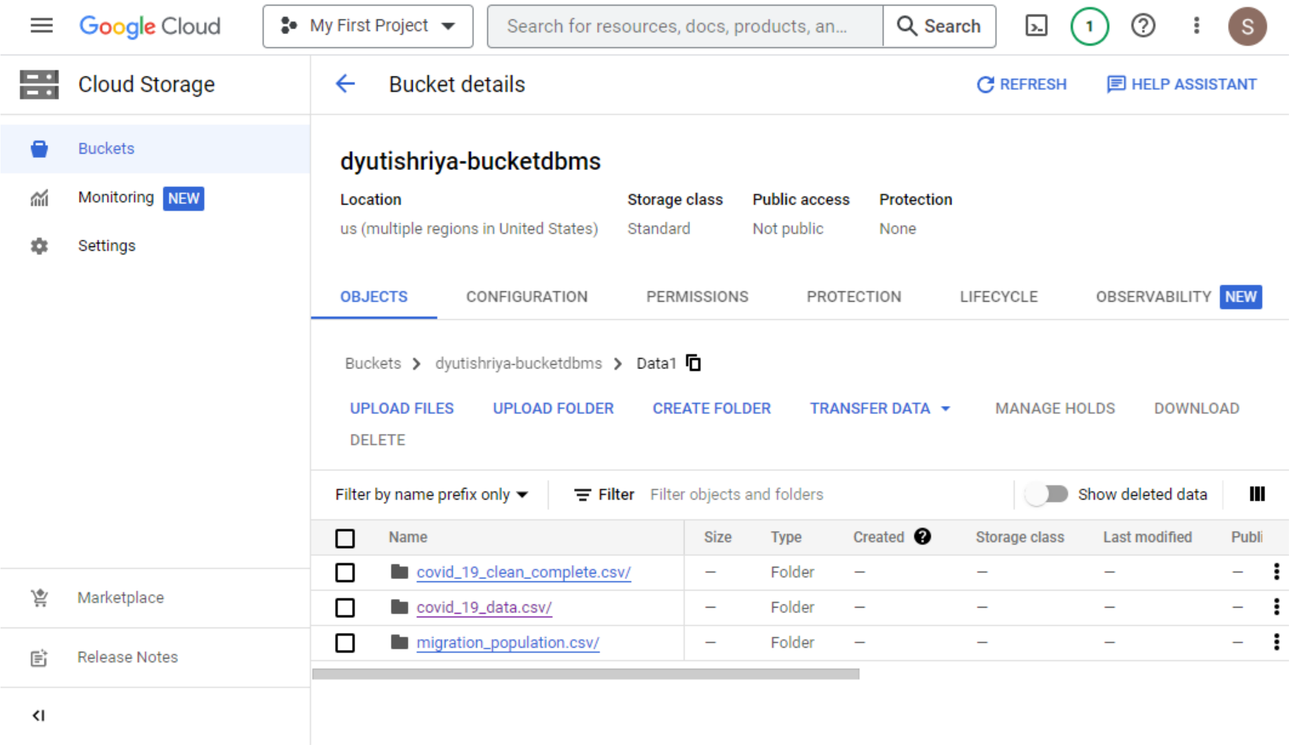




- Comparing the plots for USA and Germany its is noticeable that USA(figure down) had more active cases(blue) than Germany(figure on top)

- **Storing data back into bucket**

Data(csv files) stored in google storage bucket where subfolder is created by hadoop to save partitioned data.



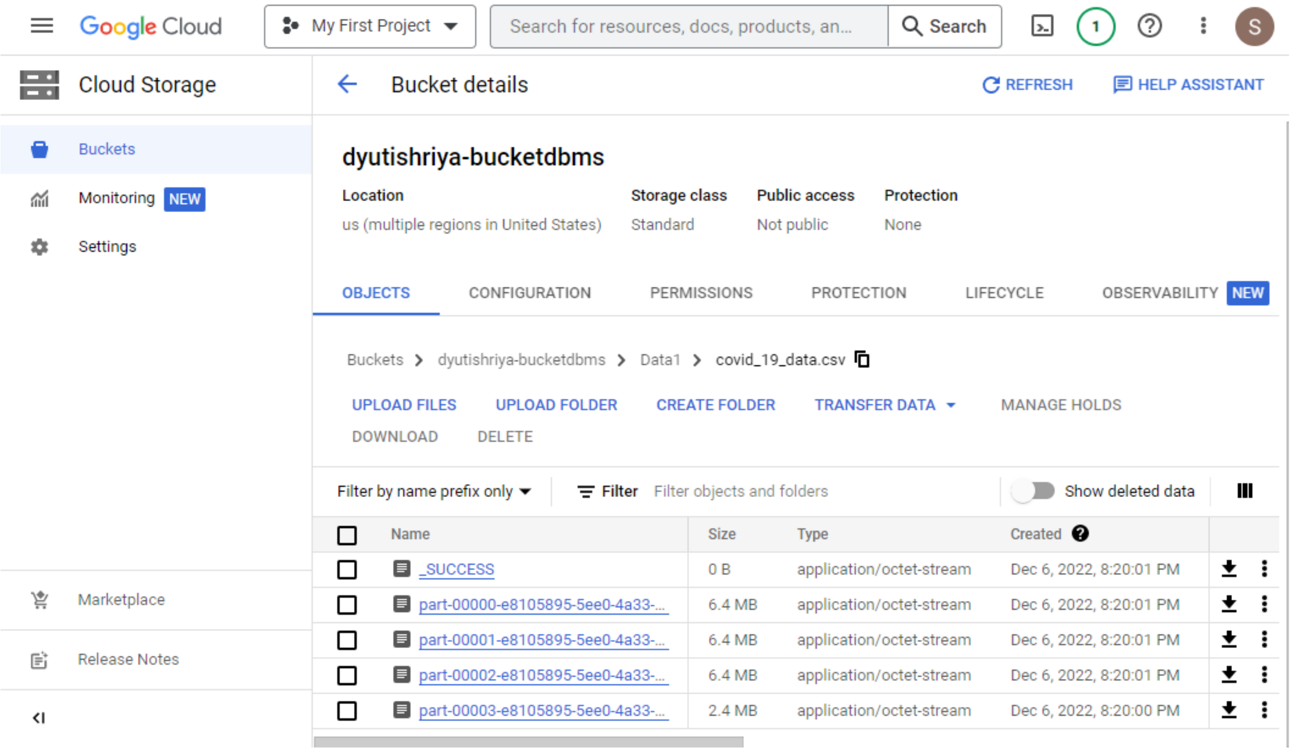
dyutishriya-bucketdbms

Location	Storage class	Public access	Protection
us (multiple regions in United States)	Standard	Not public	None

Buckets > dyutishriya-bucketdbms > Data1

	Name	Size	Type	Created	Storage class	Last modified	Publi
□	covid_19_clean_complete.csv/	—	Folder	—	—	—	—
□	covid_19_data.csv/	—	Folder	—	—	—	—
□	migration_population.csv/	—	Folder	—	—	—	—

- **Data is partitioned by hadoop into smaller chunks and saved**



dyutishriya-bucketdbms

Location	Storage class	Public access	Protection
us (multiple regions in United States)	Standard	Not public	None

Buckets > dyutishriya-bucketdbms > Data1 > covid_19_data.csv

	Name	Size	Type	Created
□	_SUCCESS	0 B	application/octet-stream	Dec 6, 2022, 8:20:01 PM
□	part-00000-e8105895-5ee0-4a33-...	6.4 MB	application/octet-stream	Dec 6, 2022, 8:20:01 PM
□	part-00001-e8105895-5ee0-4a33-...	6.4 MB	application/octet-stream	Dec 6, 2022, 8:20:01 PM
□	part-00002-e8105895-5ee0-4a33-...	6.4 MB	application/octet-stream	Dec 6, 2022, 8:20:01 PM
□	part-00003-e8105895-5ee0-4a33-...	2.4 MB	application/octet-stream	Dec 6, 2022, 8:20:00 PM

REFERENCES

- https://media.istockphoto.com/id/1215768524/vector/all-the-world-lock-down-and-stay-at-home-with-cross-line-lock-down-and-physical-distancing.jpg?s=612x612&w=0&k=20&c=IMUtLnhL9T4Du9ncS5osxslfWGG9VGNMApOyY-qG0tY=
- https://www.google.com/url?sa=i&url=https%3A%2F%2Fdatafloq.com%2Fread%2Feverything-you-need-to-know-about-big-data-2020%2F&psig=AOvVaw1nNAfqlgVI2UFsb8RPO47v&ust=1670017670722000&source=images&cd=vfe&ved=0CA8QjRxqFwoTCLjkreKy2fsCFQAAAAAdAAAAABAE
- https://bigdataprogrammers.com/load-csv-file-in-hive/