anaCovid Data Analysis

Abstract :

The project is to perform data analysis on Covid data across the globe and extract the meaningful information from the dataset which will help in taking quick and informed decision so that will help in preventing further spread of Covid virus.

Problem Statement :

1.Clean and transform data for processing

2. ETL operations on dataset

3. Storing the modified data in Azure

4.Display total cases ,new cases ,recovered cases and deaths.

5.Which country in Distinct WHO region highest cases in till date.

6. Total no of confirmed cases over between a certain date.

7.Create a geographic Map Based on total cases .

8.Date of first confirmed case in a particular region.

9.Date on which max no. cases were reported in a country.

10.Line chart showing total cases, deaths &recoveries of a particular country .

11.Bar Chart showing cases, death and recoveries reported per day/week reported.

12.Display the highest recovery rate of each month .

13.No. of active cases vs critical cases of a country.

14.Average active cases vs recoveries recorded in a day .

Technology

1. Amazon S3 Bucket- store the raw data
2. Spark- for ETL transformation and visualization(PySpark/Spark Scala/Spark SQL)
3. Amazon Redshift- for storing the processed data

Script editor 6

Redshift 8

Dynamodb 7

Glue 5,9

Data catalog (crawlers) 4