**Topic:** Covid-19 Data Lake analysis

**Team Members:** Ankitha Yenna , Sowmya Deepthi Madineni , Tharani Kumaresan, Veda Kalyan Nalluri

**Business Problem:**

COVID-19 is a viral infection caused by the SARS-CoV-2 virus. Since 2020, it has been the most dangerous word on the world. Using AWS, we can provide data on COVID-19, analyse it, and assist all researchers in taking preventive measures.

**Data set:**

The Data has been taken from Open Data Registry for Amazon Web Services. This data set contains a wealth of information on gathering resources such as the number of hospital beds and ventilators available. We can also forecast the main COVID-19 hotspots and trends using this data. This also includes information on the total number of confirmed COVID-19 cases, data on COVID-19 cases at the state and county levels in the United States, and the total number of tests performed daily. This data allows us to forecast the total number of deaths occurred and helps us to take the appropriate actions.

**Exploratory Data Analysis:**

It is used to detect any errors, outliers as well as to understand different patterns in the data using some statistical graphs and other visualization techniques. It allows us to understand the data better before making any assumptions. Here we used covid19 world cases death dataset for analysis. In Jupyter notebook, we import libraries such as numpy, matplotlib and loaded our dataset. We checked for null values, missing values, data types, asking some analytical questions and have done visualizations. These Visualizations are present in the following link <https://github.com/ankithay/BigDataproject/blob/main/project_deliverables.ipynb>

**Data Preparation:**

Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step that often involves reformatting data, making corrections to data, and the combining of data sets to enrich data. In the Data Preparation we remove the Irrelevant and Redundant data by cleaning the dataset. Our dataset contains more than 40 percent missing values, we have removed it by dropping the unnecessary columns. To clean the remaining data we have performed some Imputation Techniques. Categorical data is imputed using most frequent values and Numerical data is imputed using mean and median. The Entire Data Preparation code is present in the following link <https://github.com/ankithay/BigDataproject/blob/main/DataPreparation.ipynb>

**Data Analysis:**

The process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recapitulate and evaluate data is known as data analysis.

In this step we have analyzed and presented graphs for Date wise trends of total deaths. Analysis results are depicted in Deliverable-3 file

**Machine Learning and Optimization:**

Initially, to make Predictions of total cases from the data we have used Linear Regression. We have used Root Mean Square to evaluate the linear regression. In this case we got very large RMS value. To reduce this RMS value and to get the best fit for our values in the dataset we have used Polynomial Regression. The Results for this is mentioned in Deliverable-3.ipynb file.

**Future Work and Comments:**

1. The dataset we had to work with was extremely large, with many features having missing values. Nearly half of the features had missing data, accounting for more than 40% of the total.
2. The features are unnecessary and redundant
3. To clean the data, we deleted irrelevant and redundant features and used imputation techniques.
4. We performed linear regression to predict the total number of cases. We also used polynomial features to increase the model's performance.
5. For future work, We would like to predict hospital bed utilisation and hospitilised patients per country