Report title: MapReduce assignment

Group name: Ninjas

In this assignment we applied the MApReduce function on the COVID-19 dataset. Below are the details information about the dataset used, the problem that we tried to solve, and our approach to implement the Mapreduce functions.

Dataset :

In this work we used COVID-19 dataset provided by the Mexican government. This dataset includes an enormous amount of data that is related to the patients' pre-conditions. Such as whether the patient has medical problems like cardiovascular disease, diabetes, chronic respiratory disease, and cancer. This dataset contains 21 unique features and 1,048,576 unique patients.

Dataset columns:

- sex: female or male
- age: of the patient.
- classification: Covid test results Values 1-3 indicate that the patient has covid in varying degrees. A score of 4 or higher indicates that the patient is not a covid carrier or that the test results are ambiguous.
- patient type: whether they were hospitalized or not hospitalized.
- pneumonia: whether the patient already has air sacs inflammation or not.
- pregnancy: whether the patient is pregnant or not.
- diabetes: whether the patient has diabetes or not.
- copd: Indicates whether the patient has Chronic obstructive pulmonary disease or not.
- asthma: whether the patient has asthma or not.

- inmsupr: whether the patient is immunosuppressed or not.
- hypertension: whether the patient has hypertension or not.
- cardiovascular: whether the patient has heart or blood vessels related disease.
- renal chronic: whether the patient has chronic renal disease or not.
- other disease: whether the patient has another disease or not.
- obesity: whether the patient is obese or not.
- tobacco: whether the patient is a tobacco user.
- usmr: Indicates whether the patient treated medical units of the first, second or third level.
- medical unit: type of institution of the National Health System that provided the care.
- intubed: whether the patient was connected to the ventilator.
- icu: Indicates whether the patient had been admitted to an Intensive Care Unit.
- death: indicates whether the patient died or recovered.

Dataset Link: COVID-19 Dataset | Kaggle

Our problem:

We wanted to find out the number of deaths on the COVID-19 dataset. So we decided to use the DATE_DIED column to do that. When the value is : 9999-99-99 that means that the patient hasn't passed away, but when the column has a specific date that means that it's the death date of the patient.

Our approach to implement MapReduce functions:

First: We deleted the missing values. Which are the values with 97 and 99, as written on the dataset description .

Second: We saved the "DATE_DIED" column as a text file.

Third: We used the **MRJob** function. Which includes the mapper and the reducer functions . The mapper function maps each “date” with its count. And the reducer function sums the values for each “date”.

Lastly: Because using the **MRJob** function gave us the sum of each date, we wanted to get the sum of all those who don’t have the ‘9999-99-99’ date. So, we created another Map function that gives us the length of the dates of the people who passed away combined.