JOMO KENYATTA UNIVERSITY

OF

AGRICULTURE AND TECHNOLOGY

TAKE A WAY CAT

COURSE: B.Sc. INFORMATION TECHNOLOGY

DEPARTMENT: INFORMATION TECHNOLGY

ICS 2404

# ADVANCED DATABASE MANAGEMENT SYSTEM

REGISTRATION MEMBERS

MBABU LOYD PETER SCT221-0480/2021

SEPTEMBER-DECEMBER 2023 SEMESTER

HADOOP

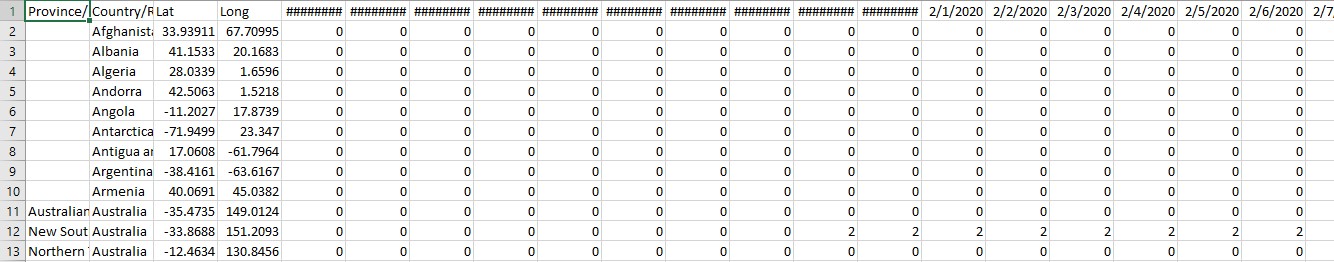
TAKE A WAY CAT WORK

NAME OF THE LECTURE: JOHN KIMOTHO

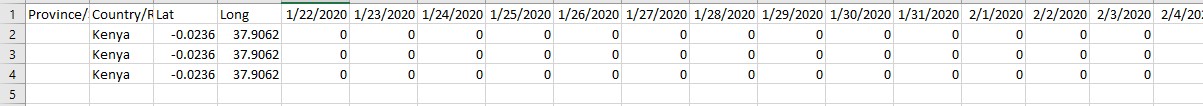
**1. How the data was assembled:** https://coronavirus.jhu.edu/ was the source from which we obtained our data.

Time\_series\_covid19\_recovered\_global.csv, Time\_series\_covid19\_deaths\_global.csv, and Time\_series\_covid19\_confirmed\_global.csv were the three excel files containing the data.

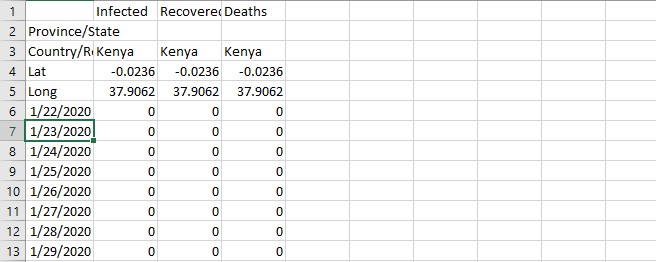
Data from every nation on the planet was included in the data sources.



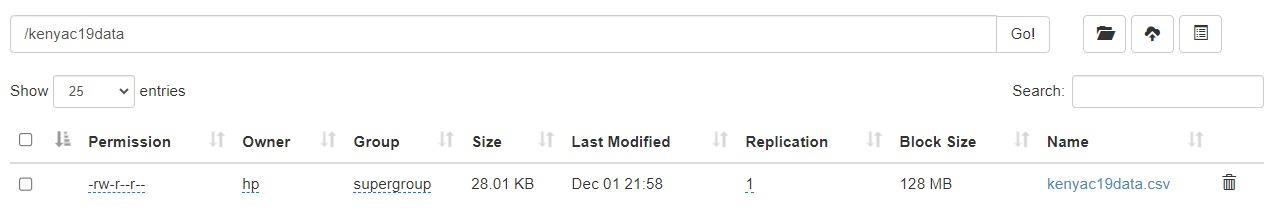
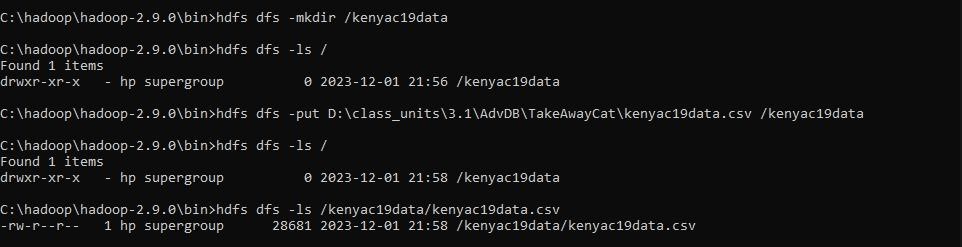
I extracted the Kenya COVID-19 data from the three csv files and pasted it into a different Excel sheet.



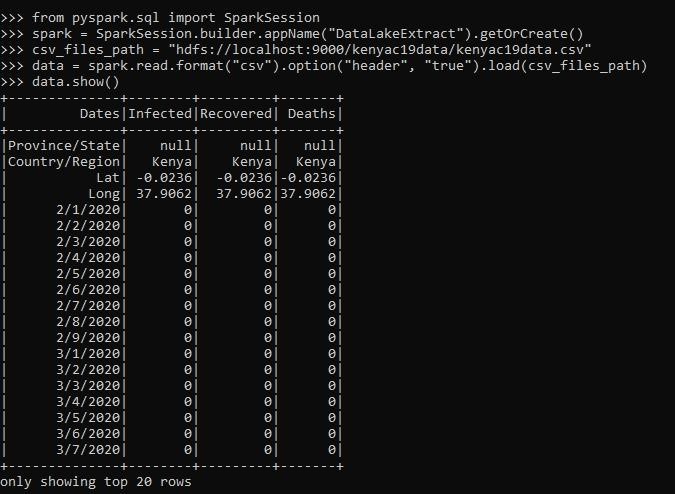
After that, I converted the Excel sheet into a CSV file by flipping the data cells and adding new column headings. I extracted the Kenya COVID-19 data from the three csv files and pasted it into a different Excel sheet.



2. **The process of ingesting data into the Hadoop data lake** involved creating a directory and inserting the prepared CSV file into it with the commands hdfs dfs -put D:\filepat and hdfs dfs -mkdir /kenyac19data.h /kenyac19data

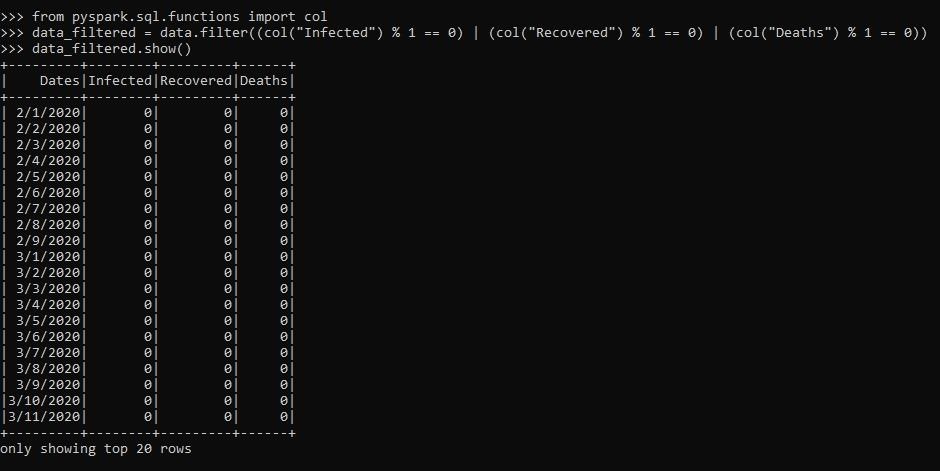


**2. How data was extracted using Pyspark**: We loaded our file from our Hadoop data lake into Pyspark by providing its path after importing the SparkSession package. The CSV file was then shown.

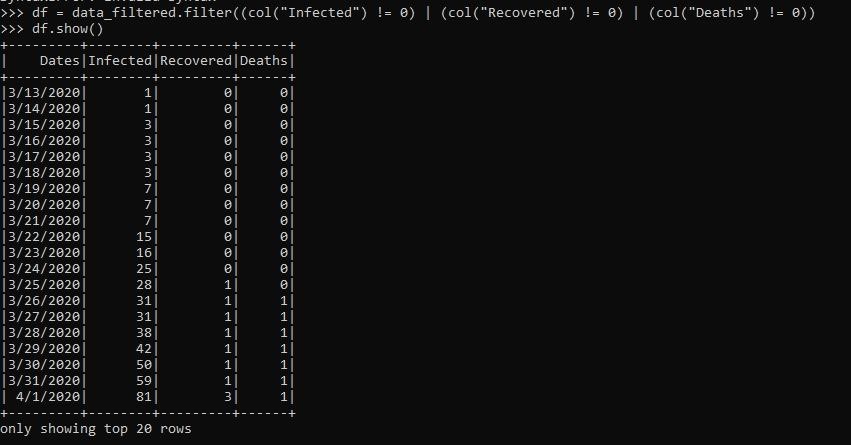


**2. The methods and tasks involved in pre-processing data.**

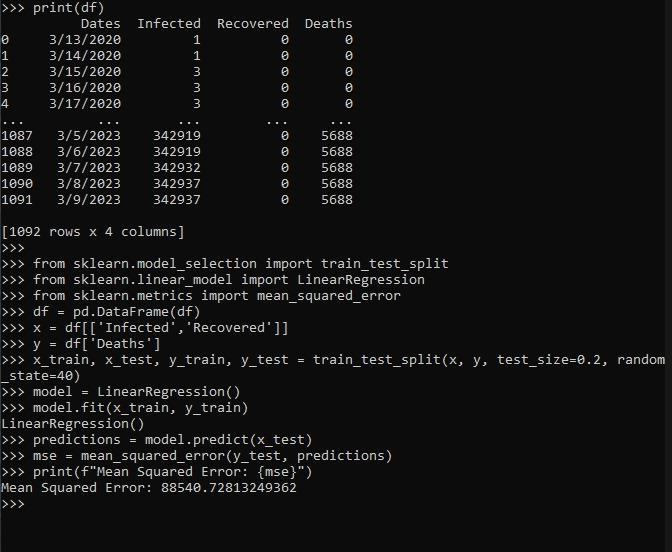
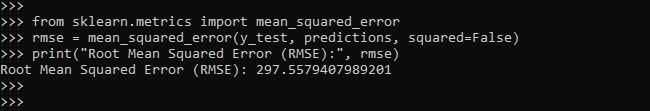
All of the rows containing null, string, and floating point values were eliminated. All that was left were easier to deal with integer values in the Infected, Recovered, and Deaths columns.



Next, eliminated every row in every column that contained a zero (0) value. I need real numbers to work with, so the rows are meaningless because all of the columns have zero values.



**2. Test Results and Interpretation**: To anticipate the number of deaths, I have employed the Mean Squared Error predictive analytics approach. The analysis revealed an MSE of 88540, which is the number of persons anticipated to die from the virus given all infections reported and the recovery rate.

**2. Results and interpretations of validations:** We used the Root Mean Squared Error approach to validate the value obtained from the Mean Squared Error technique. The result we obtained was √88540.73 = 297.56, which is the precise root of the value we discovered in the MSE.

**2. Possible uses for the interpreted findings:**

**Public Awareness:** The public can become more aware of the gravity of the issue by being informed about expected death rates based on infection levels. By encouraging adherence to vaccination schedules and preventive measures, this knowledge may help slow the virus's spread.

**b. Mitigation Strategies:** Targeted actions can be implemented in high-risk locations or among vulnerable populations, such as the elderly or those with preexisting medical issues, in response to knowledge of the estimated death toll.

**c. Making Policy:** Forecasts help policymakers make decisions concerning lockdowns, social distancing measures, travel restrictions, and vaccination campaigns by providing information about the virus's possible effects. It facilitates the development of a more focused and successful reaction.

**Data Visualization:** To see the model, I imported the matplotlib module for Python.