**Covid 19 Data analysis**

1. **Load the covid 19 data in to a spark dataframe (country\_wise\_latest.csv) with the correct schema definition**

pip install pyspark

spark = SparkSession.builder.appName("covid19\_analysis").getOrCreate()

file\_path = "mavricbdhoct01233/home/country\_wise\_latest.csv"

covid\_data\_df.show()

hdfs DFs -put country\_wise\_latest.csv  
hdfs DFs -ls

 covidDF=spark.read.option("header",True).option("inferschema",True).csv("country\_wise\_latest.csv")

covidDF.printSchema()

 covidDF.show()

 covidDF.createOrReplaceTempView("covid")

1. **The are some column names which are long, contains special characters, spaces etc. Rename all such column names accordingly. Example Country/Region → country New cases → New\_cases etc**

import pandas as pd

column\_name\_mapping = {

'Estimated Onset Date': 'Estimated',

'Contracted from which Patient (Suspected)': 'contracted',

}

df.rename(columns=column\_name\_mapping, inplace=True)

print(df.head())

1. **Count and check if there any null values in any of the columns**

null\_counts = df.isnull().sum()

any\_null\_values = df.isnull().any().any()

print("Null counts in each column:")

print(null\_counts)

print("\nAre there any null values in any column?")

print(any\_null\_values)

1. **What are the top 10 countries under the WHO region with covid 19 Confirmed cases**

from pyspark.sql import SparkSession

from pyspark.sql.window import Window

from pyspark.sql.functions import rank

spark = SparkSession.builder.appName("CovidAnalysis").getOrCreate()

window\_spec = Window.partitionBy("WHO\_Region").orderBy(covid\_data["ConfirmedCases"].desc())

ranked\_covid\_data = covid\_data.withColumn("rank", rank().over(window\_spec))

top\_10\_by\_region = ranked\_covid\_data.filter(col("rank") <= 10)

top\_10\_by\_region.show()

1. **What are the bottom 10 countries under the WHO region with covid 19 Confirmed cases**

spark = SparkSession.builder.appName("CovidAnalysis").getOrCreate()

bottom\_10\_countries = covid\_data\_with\_rank.filter(F.col("rank") <= 10)

1. **What are the total number of countries/ total no. of WHO regions and also list the various WHO regions**

from pyspark.sql import SparkSession

spark = SparkSession.builder.appName("WHORegions").getOrCreate()

countries\_df = spark.read.csv('path/to/your/data', header=True, inferSchema=True)

total\_countries = countries\_df.select(countDistinct('Country')).collect()[0][0]

print(f"Total number of countries: {total\_countries}")

total\_regions = countries\_df.select(countDistinct('WHO\_Region')).collect()[0][0]

print(f"Total number of WHO regions: {total\_regions}")

who\_regions = countries\_df.select('WHO\_Region').distinct().rdd.map(lambda x: x[0]).collect()

print("List of WHO regions:", who\_regions)