**Pyspark Assignment: Task 1**

**Covid 19 Data analysis**

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

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.createOrReplaceTempView("covid")

covidDF.printSchema()

covidDF.show()

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**2.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**

covidDF.withColumnRenamed("Deathsper100 Recovered","Deathsper100\_Recovered").show(truncate=False)

A screenshot of a computer

Description automatically generated

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

import pyspark.sql.functions as F

df\_agg = df.agg(\*[F.count(F.when(F.isnull(c), c)).alias(c) for c in df.columns])

df\_agg.show()

A screenshot of a computer

Description automatically generated

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

spark.sql( "SELECT country,WHO region,Confirmed FROM covid").show()

spark.sql( "SELECT Country,WHORegion,sum(Confirmed) FROM covid GROUP BY Country, WHORegion").show()

A screenshot of a computer

Description automatically generated

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

spark.sql( "SELECT country,WHO region,Confirmed FROM covid").show()

spark.sql( "SELECT Country,WHORegion,sum(Confirmed) FROM covid GROUP BY Country, WHORegion order by 3 asc").show(10)

A screenshot of a computer

Description automatically generated

**6. What are the total number of countries/ total no. of WHO regions and**

**also list the various WHO regions**

spark.sql("SELECT count(distinct Country), count(distinct WHORegion) from covid").show()

spark.sql("SELECT distinct WHORegion from covid").show()

A screenshot of a computer

Description automatically generated

spark.sql("SELECT count(distinct Country), count(distinct WHORegion),WHORegion from covid group by WHORegion").show()

A screenshot of a computer

Description automatically generated