**Covid 19 Data analysis**

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

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

covid.printSchema()

covid.show()

A screenshot of a computer

Description automatically generated

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**

covid.createOrReplaceTempView("covid1")

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

A close-up of a document

Description automatically generated

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

import pyspark.sql.functions as F

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

covid\_agg.show()

A black and white text

Description automatically generated

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

spark.sql( "SELECT Country,WHORegion,sum(Confirmed) FROM covid1 GROUP BY Country, WHORegion order by 3 desc").show(10)

A close-up of a computer screen

Description automatically generated

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

spark.sql( "SELECT Country,WHORegion,sum(Confirmed) FROM covid1 GROUP BY Country, WHORegion order by 3 asc").show(10)

A screenshot of a computer

Description automatically generated

1. 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),WHORegion from covid1 group by WHORegion").show()

A close-up of a computer screen

Description automatically generated