Course: Big Data

*Lab 05*

**PySpark - DataFrame**

## Question 1:

Given a tsv file [WHO-COVID-19-20210601-213841.tsv](https://drive.google.com/file/d/1TG6orBmU74s1_Z3NDsyntRb9-OAHIuy_/view?usp=sharing) which is corresponding to the [WHO Coronavirus (COVID-19) Dashboard](https://covid19.who.int/table).

Students are required to create a folder, named **lab05**, in **/content** directory of Google Colab and then copy the tsv to **/content/lab05/input/**

Take a screenshot to show your work.

|  |
| --- |
| *My screenshot* |

## Question 2:

Write a PySpark program, located in **ASEANCaseCount.py**, using DataFrames to

* to count the number of cumulative total cases among ASEAN countries (*South-East Asia Region in the given data table*)
* to find the country with the maximum number of cumulative total cases among ASEAN countries.
* to find the top 3 countries with the lowest number of cumulative cases among ASEAN countries.
* Insert your source code into the table below.

|  |
| --- |
| from pyspark.sql import SparkSession  from pyspark.sql.functions import col, sum as spark\_sum, regexp\_replace  from pyspark.sql.types import StructType, StructField, StringType, DoubleType  spark = SparkSession.builder.appName("ASEAN COVID-19 Case Count").getOrCreate()  file\_path = "/content/lab05/input/WHO-COVID-19-20210601-213841.tsv"  schema = StructType([      StructField("Name", StringType(), True),      StructField("WHO Region", StringType(), True),      StructField("Cases - cumulative total", StringType(), True),      StructField("Cases - cumulative total per 100000 population", StringType(), True),      StructField("Cases - newly reported in last 7 days", StringType(), True),      StructField("Cases - newly reported in last 7 days per 100000 population", StringType(), True),      StructField("Cases - newly reported in last 24 hours", StringType(), True),      StructField("Deaths - cumulative total", StringType(), True),      StructField("Deaths - cumulative total per 100000 population", StringType(), True),      StructField("Deaths - newly reported in last 7 days", StringType(), True),      StructField("Deaths - newly reported in last 7 days per 100000 population", StringType(), True),      StructField("Deaths - newly reported in last 24 hours", StringType(), True),      StructField("Transmission Classification", StringType(), True)  ])  df = spark.read.csv(file\_path, sep="\t", header=True, schema=schema)  df\_cleaned = df.withColumn("Cases - cumulative total",                             regexp\_replace(col("Cases - cumulative total"), ",", "").cast(DoubleType()))  asean\_countries = [ "Indonesia", "Democratic People's Republic of Korea",                      "Myanmar", "Thailand", "India", "Bangladesh", "Nepal", "Sri Lanka", "Maldives", "Timor-Leste", "Bhutan"]  asean\_df = df\_cleaned.filter((col("WHO Region") == "South-East Asia") & (col("Name").isin(asean\_countries)))  total\_cases\_row = asean\_df.agg(spark\_sum(col("Cases - cumulative total")).alias("Total Cases")).collect()[0]  total\_cases = total\_cases\_row["Total Cases"] if total\_cases\_row["Total Cases"] is not None else 0  print(f"Total cumulative COVID-19 cases in ASEAN countries: {int(total\_cases)}")  max\_case\_country = asean\_df.orderBy(col("Cases - cumulative total").desc()).first()  if max\_case\_country:      print(f"Country with the highest cases: {max\_case\_country['Name']} - {int(max\_case\_country['Cases - cumulative total'])} cases")  else:      print("No data found for highest cases.")  lowest\_cases\_df = asean\_df.orderBy(col("Cases - cumulative total").asc()).select("Name", "Cases - cumulative total").limit(3)  print("Top 3 countries with the lowest cumulative cases in ASEAN:")  for row in lowest\_cases\_df.collect():      print(f"{row['Name']} - {int(row['Cases - cumulative total'])} cases")  spark.stop() |

* Take a screenshot of the terminal to visualize the program result.

|  |
| --- |
| *My screenshot* |