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.

|  |
| --- |
|  |

## 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 udf, col  from pyspark.sql.types import FloatType  import re  from tabulate import tabulate  spark = SparkSession.builder.appName("ASEANCaseCount").getOrCreate()  data = spark.read.csv("/content/lab05/input/WHO-COVID-19-20210601-213841.tsv", sep="\t", header=True, inferSchema=True)  asean\_data = data.filter(data["WHO Region"] == "South-East Asia")  remove\_commas = udf(lambda x: float(re.sub('[,.]', '', x)), FloatType())  asean\_data = asean\_data.withColumn("Cases - cumulative total", remove\_commas("Cases - cumulative total"))  print("\n==================== Question 1 ====================\n")  print("Number of cumulative total cases among ASEAN countries:", asean\_data.groupby('Name').sum("Cases - cumulative total").collect()[0][1])  print("\n==================== Question 2 ====================\n")  max\_cases\_country = asean\_data.orderBy(col("Cases - cumulative total").desc()).first()  print("Country with the maximum number of cumulative total cases among ASEAN countries:", max\_cases\_country[0])  print("Maximum Cumulative Cases:", max\_cases\_country[2])  print("\n==================== Question 3 ====================\n")  lowest\_cases\_countries = asean\_data.orderBy(col("Cases - cumulative total")).limit(3).collect()  print("Top 3 countries with the lowest number of cumulative cases among ASEAN countries:")  print(tabulate([[country[0], country[2]] for country in lowest\_cases\_countries], headers=["Country", "Cases - cumulative total"], tablefmt="grid"))  spark.stop() |

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

|  |
| --- |
|  |

## Submission Notice

* Export your answer file as pdf
* Rename the pdf following the format:

**lab05\_<student number>\_HoTen.pdf**

E.g. lab05\_123456\_NguyenThanhAn.pdf

*If you have not been assigned a student number yet, then use 123456 instead.*

* Careless mistakes in filename, format, question order, etc. are not accepted (0 pts).