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.

|  |
| --- |
| *Your screenshot goes here* |

## 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    def filter\_south\_east\_asia(line):  parts = line.split('\t')  region = parts[1].strip()  return region == "South-East Asia"    def parse\_and\_filter(line):  parts = line.split('\t')  country = parts[0]  cumulative\_total\_str = parts[2].replace(',','')  cumulative\_total = float(cumulative\_total\_str)  return country, cumulative\_total    if \_\_name\_\_ == "\_\_main\_\_":  # Create instance of SparkSession  spark = SparkSession.builder.appName("ASEANCaseCount").getOrCreate()    input\_path = 'hdfs://localhost:9000/user/hoang/lab04/input/WHO-COVID-19-20210601-213841.tsv'    input\_rdd = spark.sparkContext.textFile(input\_path)    filtered\_rdd = input\_rdd.filter(filter\_south\_east\_asia)    asean\_rdd = filtered\_rdd.map(parse\_and\_filter)  print(asean\_rdd.count())    cumulative\_total = asean\_rdd.map(lambda x: x[1]).reduce(lambda x, y: x + y)    print("Cumulative total cases among ASEAN countries in South-East Asia Region:",  cumulative\_total)    spark.stop() |

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

|  |
| --- |
| *Your screenshot goes here* |

## 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).