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  from pyspark.sql.functions import col, max  spark = SparkSession.builder.appName('ASEANCaseCount').getOrCreate()  data = spark.read.csv('/content/drive/MyDrive/Colab\_Notebooks/BigData/lab05/input/WHO-COVID-19-20210601-213841.tsv', header=True, sep='\t')  asean\_countries = ['Brunei Darussalam', 'Cambodia', 'Indonesia', 'Lao People\'s Democratic Republic', 'Malaysia', 'Myanmar', 'Philippines', 'Singapore', 'Thailand', 'Viet Nam']  data = data.filter((data['WHO Region'] == 'South-East Asia') & (data['Name'].isin(asean\_countries)))  data = data.withColumn('Cases - cumulative total', data['Cases - cumulative total'].cast('integer'))  total\_cases = data.groupBy().sum('Cases - cumulative total').collect()[0][0]  print(f'Total cases in ASEAN countries: {total\_cases}')  max\_cases\_country = data.orderBy(data['Name'].desc()).first()['Name']  print(f'Country with the maximum number of cases: {max\_cases\_country}')  lowest\_cases\_countries = data.orderBy(data['Cases - cumulative total']).select('Name').limit(3).collect()  lowest\_cases\_countries = [row['Name'] for row in lowest\_cases\_countries]  print(f'Top 3 countries with the lowest number of cases: {lowest\_cases\_countries}')  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).