Course: Big Data

Lab 05

PySpark - DataFrame

Question 1:

Given a tsv file [WHO-COVID-19-20210601-213841.tsv] which is corresponding to the [WHO Coronavirus (COVID-19) Dashboard].

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:

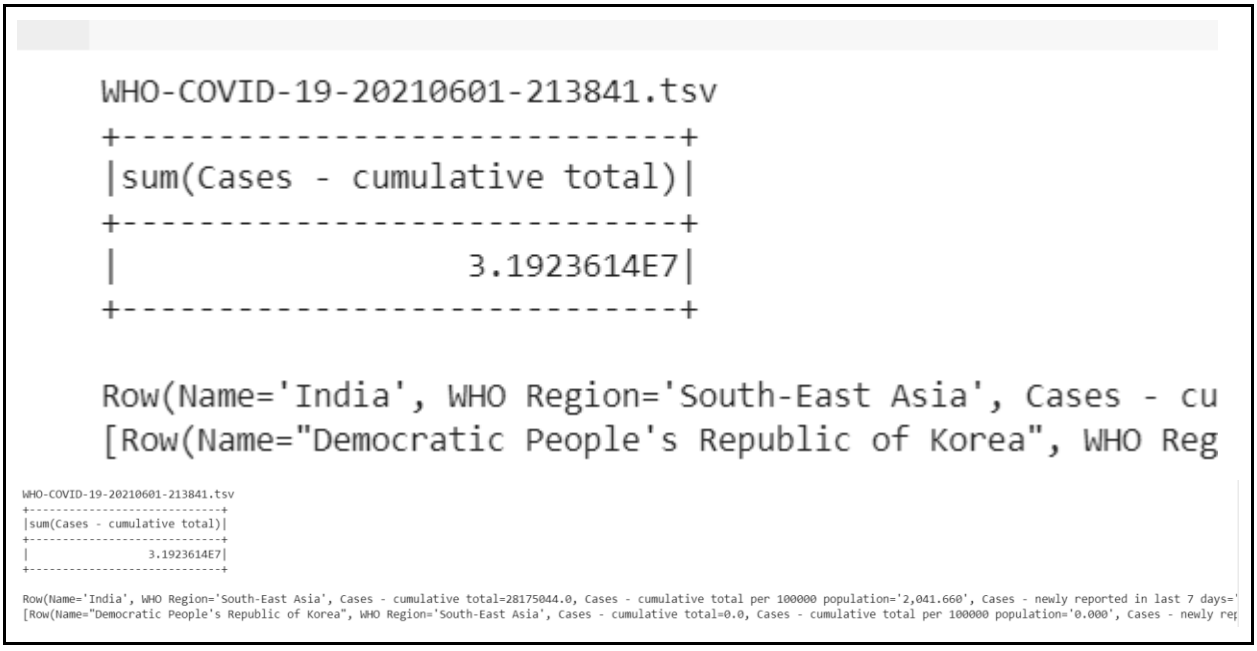
Write a PySpark program, located in **ASEANCaseCount.py**, using DataFrames to

- 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 col, udf, sum
from pyspark.sql.types import FloatType
# Create SparkSession
spark = SparkSession.builder \
    .appName("MMDS-Lab05") \
    .getOrCreate()
# Define paths and constants
DATA_PATH = '/content/lab05/input/'
TSV_FILE = '/content/WHO-COVID-19-20210601-213841.tsv'
# Create the directory lab05 in /content
!mkdir -p /content/lab05/input
# Copy the tsv file to /content/lab05/input/
!cp {TSV_FILE} /content/lab05/input/
# Verify that the file has been copied
!ls /content/lab05/input/

# Define separator character
SEPARATED_CHAR = '\t'
ASEAN_COUNTRIES = ['South-East Asia']

# Read data
case_string_2_list = udf(lambda s: float(s.replace(',', '')),
FloatType())
data = spark.read.csv(DATA_PATH, sep=SEPARATED_CHAR, header=True)\
            .withColumn('Cases - cumulative total',
case_string_2_list(col('Cases - cumulative total')))

# Filter ASEAN countries
asean_countries = data.where(col('WHO Region') == 'South-East Asia')

# Task 1: Count the number of cumulative total cases among ASEAN
countries
asean_countries.select(sum(col('Cases - cumulative total'))).show()

# Task 2: Find the country with the maximum number of cumulative total
cases among ASEAN countries.
print(asean_countries.orderBy('Cases - cumulative total',
ascending=False).first())

# Task 3: Find the top 3 countries with the lowest number of cumulative
cases among ASEAN countries.
print(asean_countries.orderBy('Cases - cumulative total',
ascending=True).take(3))
```

- 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>_<full name>.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).