

# Course: Big Data

## Lab 05

### PySpark - DataFrame

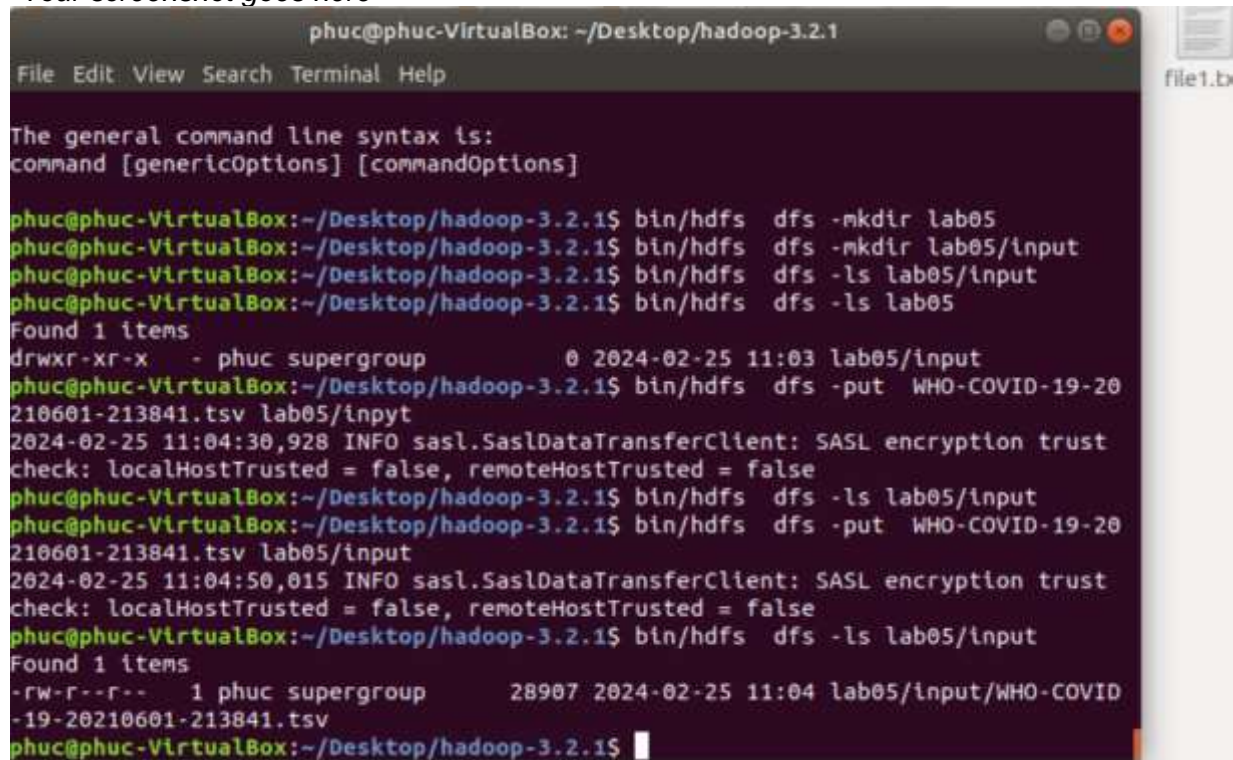
#### Question 1:

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.

Your screenshot goes here



```
phuc@phuc-VirtualBox: ~/Desktop/hadoop-3.2.1
File Edit View Search Terminal Help

The general command line syntax is:
command [genericOptions] [commandOptions]

phuc@phuc-VirtualBox:~/Desktop/hadoop-3.2.1$ bin/hdfs dfs -mkdir lab05
phuc@phuc-VirtualBox:~/Desktop/hadoop-3.2.1$ bin/hdfs dfs -mkdir lab05/input
phuc@phuc-VirtualBox:~/Desktop/hadoop-3.2.1$ bin/hdfs dfs -ls lab05/input
phuc@phuc-VirtualBox:~/Desktop/hadoop-3.2.1$ bin/hdfs dfs -ls lab05
Found 1 items
drwxr-xr-x - phuc supergroup          0 2024-02-25 11:03 lab05/input
phuc@phuc-VirtualBox:~/Desktop/hadoop-3.2.1$ bin/hdfs dfs -put WHO-COVID-19-20
210601-213841.tsv lab05/inpyt
2024-02-25 11:04:30,928 INFO sasl.SaslDataTransferClient: SASL encryption trust
check: localhostTrusted = false, remoteHostTrusted = false
phuc@phuc-VirtualBox:~/Desktop/hadoop-3.2.1$ bin/hdfs dfs -ls lab05/input
phuc@phuc-VirtualBox:~/Desktop/hadoop-3.2.1$ bin/hdfs dfs -put WHO-COVID-19-20
210601-213841.tsv lab05/input
2024-02-25 11:04:50,015 INFO sasl.SaslDataTransferClient: SASL encryption trust
check: localhostTrusted = false, remoteHostTrusted = false
phuc@phuc-VirtualBox:~/Desktop/hadoop-3.2.1$ bin/hdfs dfs -ls lab05/input
Found 1 items
-rw-r--r--  1 phuc supergroup      28907 2024-02-25 11:04 lab05/input/WHO-COVID
-19-20210601-213841.tsv
phuc@phuc-VirtualBox:~/Desktop/hadoop-3.2.1$
```

## 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.

```
# -*- coding: utf-8 -*-
"""lab05_521H0509_BigData

Automatically generated by Colaboratory.

Original file is located at
    https://colab.research.google.com/drive/1WnUZm1melB4_psqacnU15iRyN98U14N9
"""

# !apt-get install openjdk-8-jdk-headless -qq > /dev/null
# !wget -q http://archive.apache.org/dist/spark/spark-3.1.1/spark-3.1.1-bin-
hadoop3.2.tgz
# # !cp drive/MyDrive/MMDS-data/spark-3.1.1-bin-hadoop3.2.tgz .
# !tar xf spark-3.1.1-bin-hadoop3.2.tgz
# !pip install -q findspark

import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-1.8.0-openjdk-amd64"
os.environ["SPARK_HOME"] = "/home/phuc/Desktop/spark-3.1.1-bin-hadoop3.2"

# os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
# os.environ["SPARK_HOME"] = "/content/spark-3.1.1-bin-hadoop3.2"

import findspark
```

```

findspark.init(os.environ["SPARK_HOME"])

from pyspark.sql import SQLContext
from pyspark import SparkContext
import pyspark as spark
from pyspark.sql.functions import expr
from pyspark.sql.functions import regexp_replace
from pyspark.sql.functions import max
from pyspark.sql.functions import sum
print(spark.__version__)
sc = SparkContext("local", "Second App")

sqlc = SQLContext(sc)
df = sqlc.read.csv('hdfs://localhost:9000/user/phuc/lab05/input/WHO-COVID-19-
20210601-213841.tsv', sep = "\t",
                  header=True, inferSchema=True)
# df = sqlc.read.csv('WHO-COVID-19-20210601-213841.tsv', sep = "\t",
#                   header=True, inferSchema=True)
df.show()

df = df.withColumn('Cases - cumulative total', expr("translate('Cases - cumulative
total', ',', ' ')").cast('int'))

df.where(df['WHO Region'] == "South-East Asia")\
  .groupBy('WHO Region')\
  .agg(sum('Cases - cumulative total').alias("number of cumulative total"))\
  .show()

max_cases = df.where(df['WHO Region'] == 'South-East Asia')\
  .agg(max('Cases - cumulative total').alias('max_cases'))\
  .collect()[0]['max_cases']

df.where(df['WHO Region'] == 'South-East Asia')\
  .where(df['Cases - cumulative total'] == max_cases)\
  .select('Name')\
  .show()

df.where(df['WHO Region'] == 'South-East Asia')\
  .sort(df['Cases - cumulative total'])\
  .select(df['Name'])\
  .show(3)

```

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

Your screenshot goes here

```
+-----+
| WHO Region|number of cumulative total|
+-----+
|South-East Asia|          31923614|
+-----+

+-----+
| Name|
+-----+
|India|
+-----+

+-----+
|          Name|
+-----+
|Democratic People...|
|          Bhutan|
|          Timor-Leste|
+-----+
only showing top 3 rows
phuc@phuc-VirtualBox:~/Downloads$
```

## 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).