Course: Big Data

Lab 04

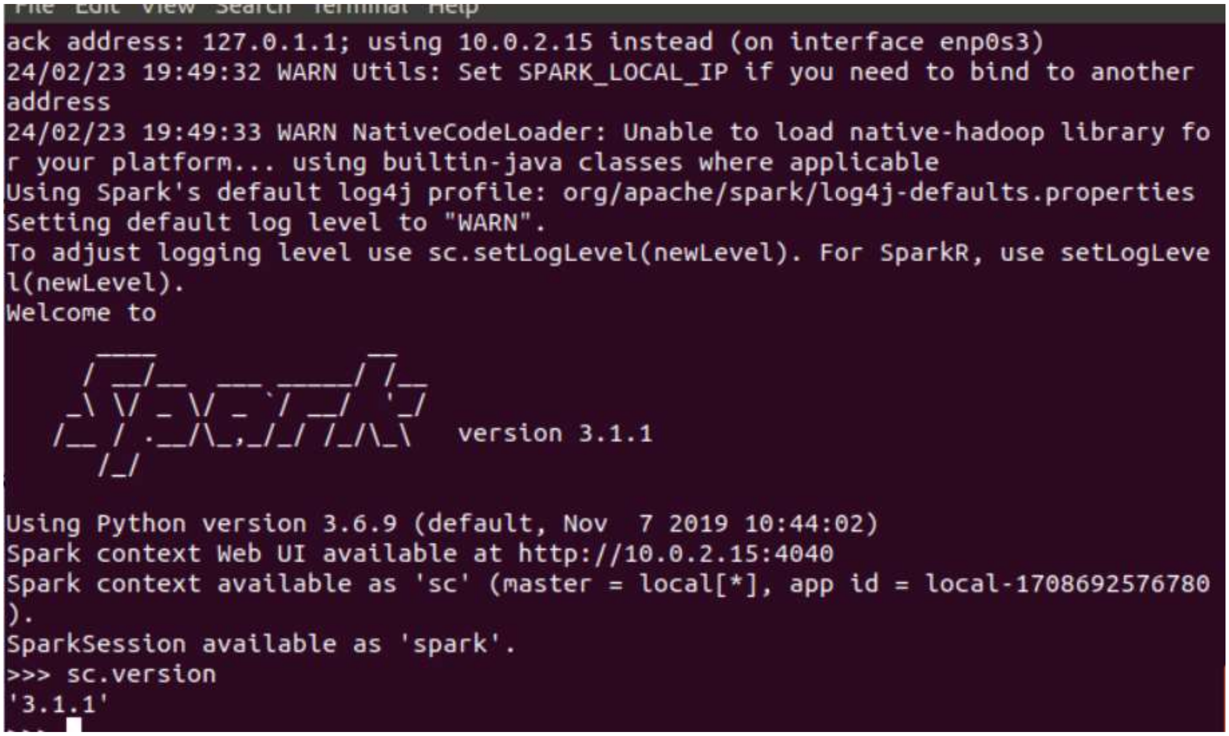
PySpark - RDD

Question 1:

Based on the tutorial of PySpark, students install PySpark in Ubuntu.

- Define the environment variable: JAVA_HOME
- Define the environment variable: SPARK_HOME
- Start the pyspark-shell and write an instruction to print down the PySpark version
- Take the screenshot and insert it into the table below.

Your screenshot goes here



```
ack address: 127.0.1.1; using 10.0.2.15 instead (on interface enp0s3)
24/02/23 19:49:32 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
24/02/23 19:49:33 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLeve
l(newLevel).
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.1.1
      /_/

Using Python version 3.6.9 (default, Nov  7 2019 10:44:02)
Spark context Web UI available at http://10.0.2.15:4040
Spark context available as 'sc' (master = local[*], app id = local-1708692576780
).
SparkSession available as 'spark'.
>>> sc.version
'3.1.1'
>>> 
```

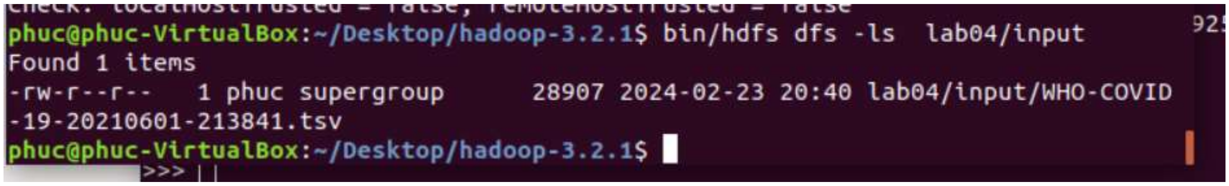
Question 2:

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 **lab04**, in HDFS and then copy the tsv to **lab04/input/**

Take a screenshot to show the content of **lab04/input/** in HDFS

Your screenshot goes here



```
check: localhostTrusted = false, remoteHostTrusted = false
phuc@phuc-VirtualBox:~/Desktop/hadoop-3.2.1$ bin/hdfs dfs -ls  lab04/input
Found 1 items
-rw-r--r--   1 phuc supergroup      28907 2024-02-23 20:40 lab04/input/WHO-COVID
-19-20210601-213841.tsv
phuc@phuc-VirtualBox:~/Desktop/hadoop-3.2.1$ 
```

Question 3:

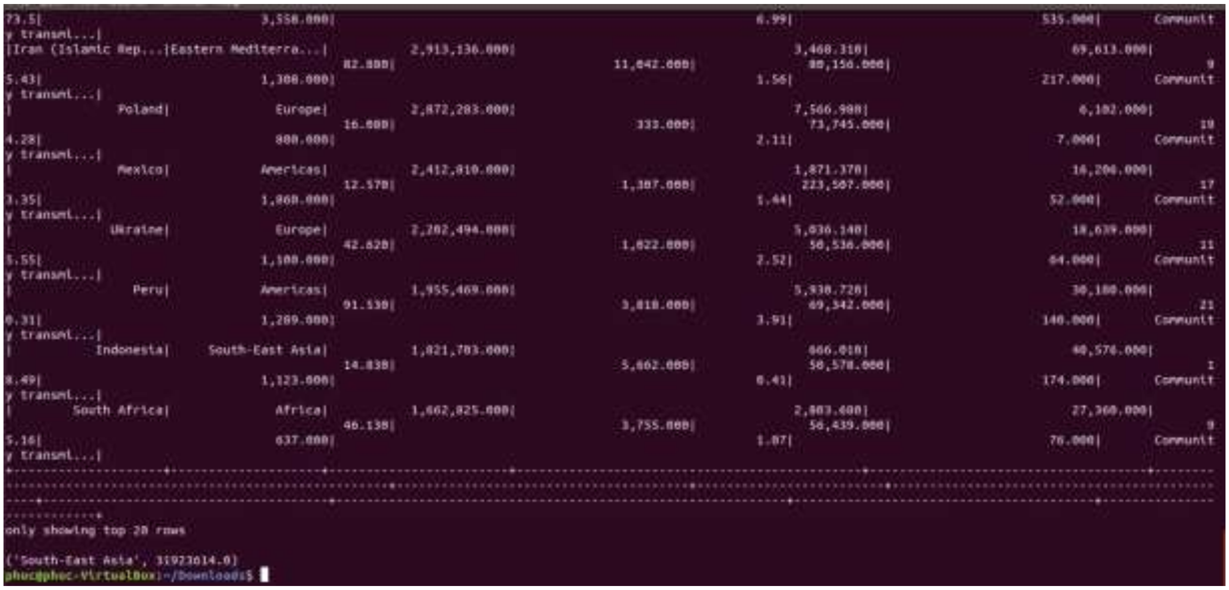
Write a PySpark program, located in **ASEANCaseCount.py**, to count the number of cumulative total cases among ASEAN countries (*South-East Asia Region in the given data table*) using RDDs.

- Insert your source code into the table below.

```
# -*- coding: utf-8 -*-
"""Lab04_521H0509

Automatically generated by Colaboratory.

Original file is located at
    https://colab.research.google.com/drive/11nwdweTCs0gT3ijTBua4unMzPD0Ke3uC
"""

# !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"

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

findspark.init()
from pyspark.sql import SQLContext
from pyspark import SparkContext
import pyspark as spark
sc = SparkContext("local", "First App")
print(spark.__version__)

sqlc = SQLContext(sc)
df = sqlc.read.csv('hdfs://localhost:9000/user/phuc/lab04/input/WHO-COVID-19-
20210601-213841.tsv',sep = "\t",
                   header=True, inferSchema=True)
df.show()

rddObj=df.rdd

# print(*rddObj.collect(), sep='\n')


result =  rddObj.filter(lambda x: x['WHO Region'] =='South-East Asia')\
                        .map(lambda x: (x['WHO Region'], float(x['Cases -
cumulative total'].replace(',', '')) ) ) \
                        .reduceByKey(lambda x, y: x + y)
print(*result.collect())
```

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

Your screenshot goes here



```
73.5|                             3,558.000|                                                         6.99|                                 535.000|       Communit
y transml...|
|Iran (Islamic Rep...|Eastern Mediterra...|           2,913,136.000|                                      3,468.310|                           69,613.000|
                                             82.800|                              11,042.000|               80,156.000|                                          9
5.43|                             1,308.000|                                                         1.56|                                 217.000|       Communit
y transml...|
|              Poland|              Europe|           2,872,283.000|                                      7,566.980|                            6,102.000|
                                             16.080|                                 333.000|               73,745.000|                                         19
4.28|                               800.000|                                                         2.11|                                   7.000|       Communit
y transml...|
|              Mexico|            Americas|           2,412,010.000|                                      1,871.370|                           16,200.000|
                                             12.570|                               1,387.000|              223,507.000|                                         17
3.35|                             1,860.000|                                                         1.44|                                  52.000|       Communit
y transml...|
|             Ukraine|              Europe|           2,282,494.000|                                      5,036.140|                           18,639.000|
                                             42.620|                               1,822.000|               50,536.000|                                         11
5.55|                             1,100.000|                                                         2.52|                                  64.000|       Communit
y transml...|
|                Peru|            Americas|           1,955,469.000|                                      5,930.720|                           30,100.000|
                                             91.530|                               3,818.000|               69,342.000|                                         21
0.31|                             1,289.000|                                                         3.91|                                 140.000|       Communit
y transml...|
|           Indonesia|     South-East Asia|           1,021,703.000|                                        666.010|                           40,576.000|
                                             14.830|                               5,662.000|               50,570.000|                                          I
8.49|                             1,123.000|                                                         0.41|                                 174.000|       Communit
y transml...|
|        South Africa|              Africa|           1,662,025.000|                                      2,003.600|                           27,360.000|
                                             46.130|                               3,755.000|               56,439.000|                                          9
5.16|                               637.000|                                                         1.07|                                  76.000|       Communit
y transml...|
+--------------------+--------------------+------------------------+-----------------------------------------------+-------------------------------------+--------
---------------------------------------------------+----------------------------------------+-------------------------+-------------------------------------------
----+--------------------------------------+-------------------------------------------------------------+----------------------------------------+---------------
------------+
only showing top 20 rows

('South-East Asia', 31923614.0)
phuc@phuc-VirtualBox:~/Downloads$ 
```

Submission Notice

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

  lab04_<student number>_HoTen.pdf

  E.g. lab04_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).