Course: Big Data

*Lab 04*

**PySpark - RDD**

## Question 1:

Based on [the tutorial of PySpark](https://colab.research.google.com/drive/1WgRu3QVcaG9McOlXbiJWhBuWdjkbuNct?usp=sharing), 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* |

## Question 2:

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 **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* |

## 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* |

## 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).