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.

|  |
| --- |
|  |

## 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

|  |
| --- |
|  |

## 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.

|  |
| --- |
|  |

* 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:

**lab04\_<student number>\_<full name>.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).