Course: Big Data

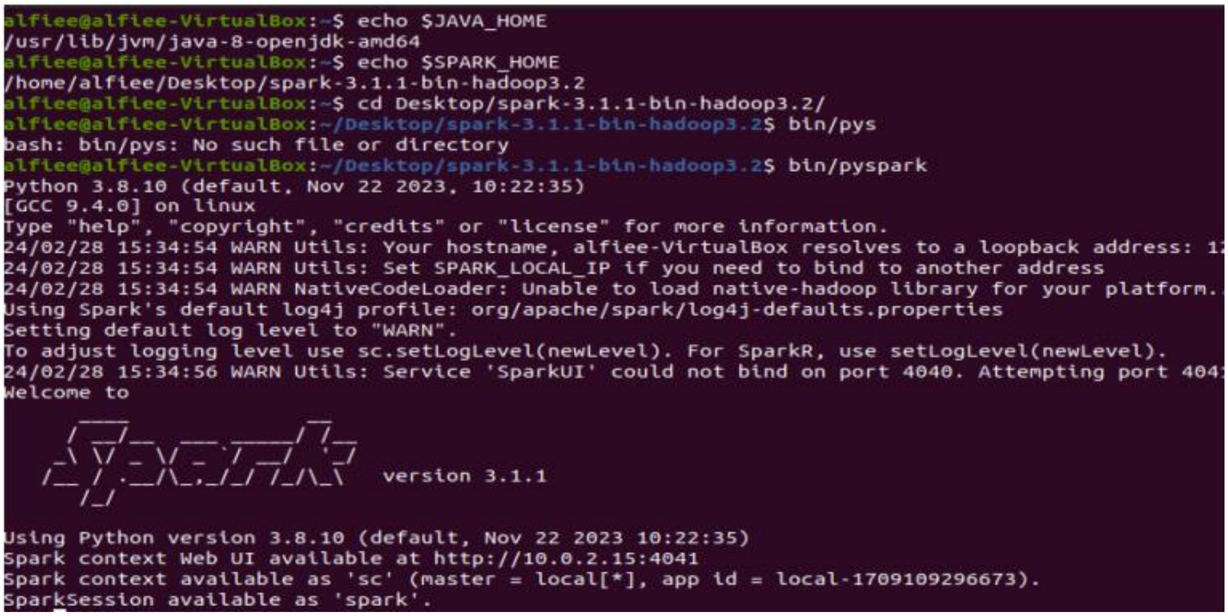
Lab 04

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.



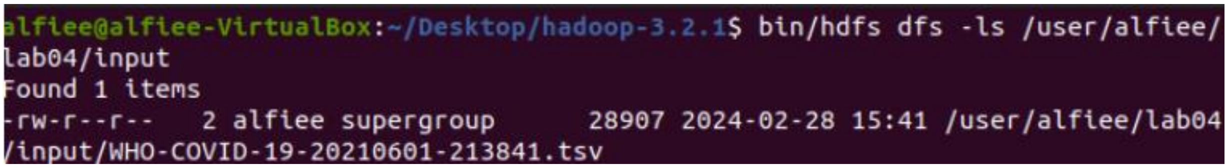
```
alfiee@alfiee-VirtualBox:~$ echo $JAVA_HOME
/usr/lib/jvm/java-8-openjdk-amd64
alfiee@alfiee-VirtualBox:~$ echo $SPARK_HOME
/home/alfiee/Desktop/spark-3.1.1-bin-hadoop3.2
alfiee@alfiee-VirtualBox:~$ cd Desktop/spark-3.1.1-bin-hadoop3.2/
alfiee@alfiee-VirtualBox:~/Desktop/spark-3.1.1-bin-hadoop3.2$ bin/pys
bash: bin/pys: No such file or directory
alfiee@alfiee-VirtualBox:~/Desktop/spark-3.1.1-bin-hadoop3.2$ bin/pyspark
Python 3.8.10 (default, Nov 22 2023, 10:22:35)
[GCC 9.4.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
24/02/28 15:34:54 WARN Utils: Your hostname, alfiee-VirtualBox resolves to a loopback address: 1
24/02/28 15:34:54 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
24/02/28 15:34:54 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform.
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 setLogLevel(newLevel).
24/02/28 15:34:56 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 404
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.1.1
      /_/

Using Python version 3.8.10 (default, Nov 22 2023 10:22:35)
Spark context Web UI available at http://10.0.2.15:4041
Spark context available as 'sc' (master = local[*], app id = local-1709109296673).
SparkSession available as 'spark'.
```

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

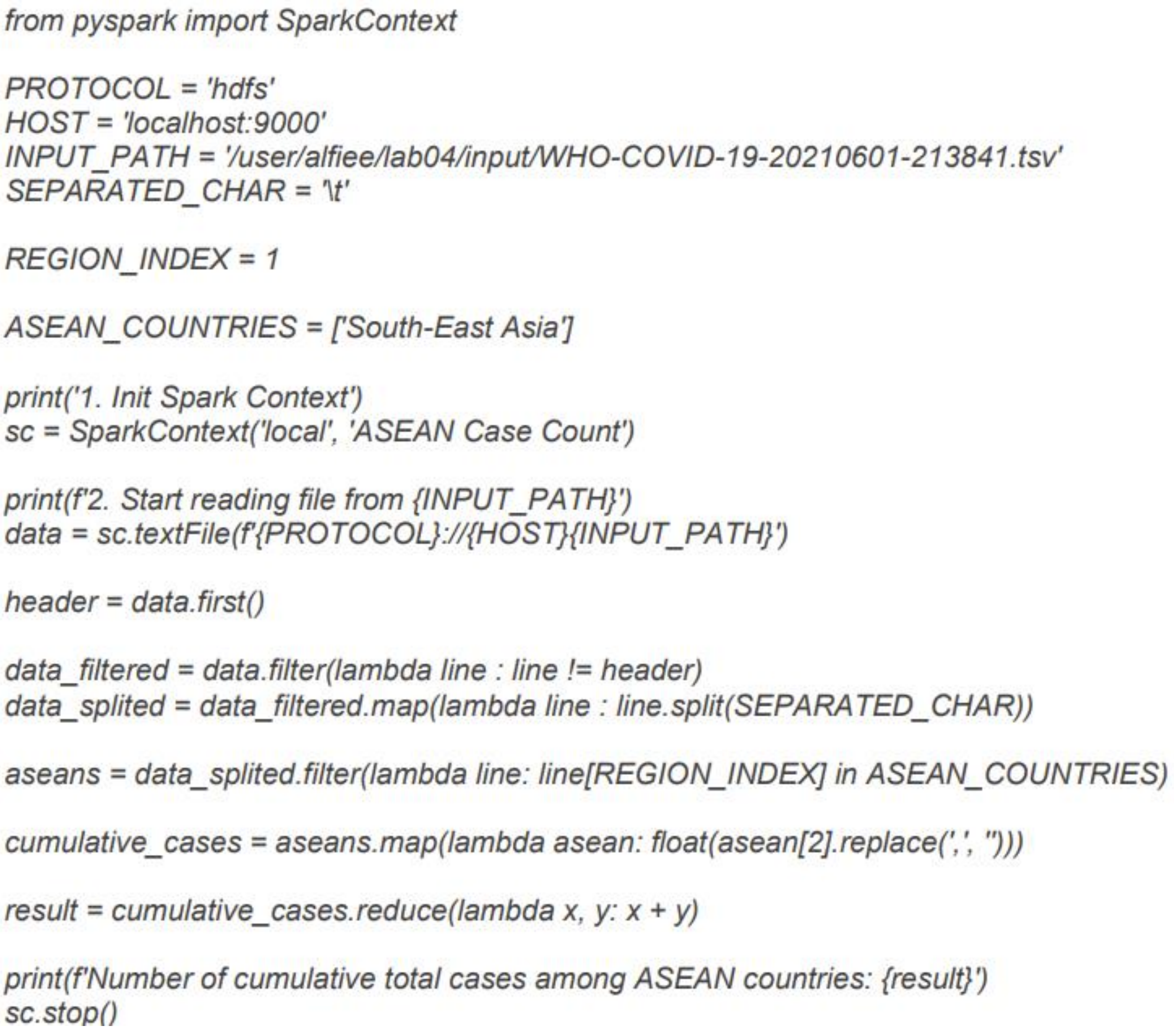


```
alfiee@alfiee-VirtualBox:~/Desktop/hadoop-3.2.1$ bin/hdfs dfs -ls /user/alfiee/
lab04/input
Found 1 items
-rw-r--r--   2 alfiee supergroup      28907 2024-02-28 15:41 /user/alfiee/lab04
/input/WHO-COVID-19-20210601-213841.tsv
```

Question 3:

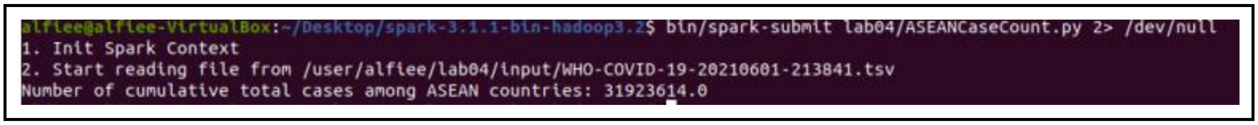
Write a PySpark program, located in **ASEANCaseCount.py**, to count the number of cumulative total cases among ASEAN countries (*South-East Asia Region in the given data table*) using RDDs.

- Insert your source code into the table below.



```
from pyspark import SparkContext

PROTOCOL = 'hdfs'
HOST = 'localhost:9000'
INPUT_PATH = '/user/alfiee/lab04/input/WHO-COVID-19-20210601-213841.tsv'
SEPARATED_CHAR = '\t'

REGION_INDEX = 1

ASEAN_COUNTRIES = ['South-East Asia']

print('1. Init Spark Context')
sc = SparkContext('local', 'ASEAN Case Count')

print(f'2. Start reading file from {INPUT_PATH}')
data = sc.textFile(f'{PROTOCOL}://{HOST}{INPUT_PATH}')

header = data.first()

data_filtered = data.filter(lambda line : line != header)
data_splited = data_filtered.map(lambda line : line.split(SEPARATED_CHAR))

aseans = data_splited.filter(lambda line: line[REGION_INDEX] in ASEAN_COUNTRIES)

cumulative_cases = aseans.map(lambda asean: float(asean[2].replace(',', '')))

result = cumulative_cases.reduce(lambda x, y: x + y)

print(f'Number of cumulative total cases among ASEAN countries: {result}')
sc.stop()
```

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



```
alfiee@alfiee-VirtualBox:~/Desktop/spark-3.1.1-bin-hadoop3.2$ bin/spark-submit lab04/ASEANCaseCount.py 2> /dev/null
1. Init Spark Context
2. Start reading file from /user/alfiee/lab04/input/WHO-COVID-19-20210601-213841.tsv
Number of cumulative total cases among ASEAN countries: 31923614.0
```

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).