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.

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

```
hohuuan@ubuntu:~/Desktop/hadoop$ bin/hdfs dfs -ls lab04/input
Found 1 items
-rw-r--r--   2 hohuuan supergroup      28907 2024-01-27 14:40 lab04/input/WHO-COVID-19-20210601-213841.tsv
```

```
hohuuan@ubuntu:~/Desktop/hadoop$ bin/hdfs dfs -cat lab04/input/WHO-COVID-19-20210601-213841.tsv
2024-01-27 14:42:43,846 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Name    WHO Region      Cases - cumulative total        Cases - cumulative total per 100000 population  Cases - newly reported in last 7 days    Cases - newly reported in last 7 days per 100000 population     Ca
ses - newly reported in last 24 hours Deaths - cumulative total       Deaths - cumulative total per 100000 population Deaths - newly reported in last 7 days  Deaths - newly reported in last 7 days per 100000 po
pulation      Deaths - newly reported in last 24 hours        Transmission Classification
Global -        170,363,852.000 2,182.386       3,383,026.000   43.337  358,691.000     3,546,870.000   45.436  76,025.000      0.974   7,370.000
United States of America        Americas        32,929,178.000  9,948.310       131,305.000     39.670  0.000   588,596.000     177.820  3,896.000       1.180   0.000   Community transmission
India   South-East Asia 28,175,044.000  2,041.660       1,226,170.000   88.850  127,510.000     331,895.000     24.050  24,664.000      1.790   2,795.000       Clusters of cases
Brazil  Americas        16,515,120.000  7,769.650       431,862.000     203.170 43,520.000      461,931.000     217.320 12,863.000      6.050   874.000 Community transmission
France  Europe  5,566,214.000   8,558.220       59,437.000      91.390  8,541.000       108,558.000     166.910 756.000 1.160   15.000  Community transmission
Turkey  Europe  5,242,911.000   6,216.470       56,424.000      66.900  6,933.000       47,405.000      56.210  1,137.000       1.350   134.000 Community transmission
Russian Federation      Europe  5,071,917.000   3,475.480       62,006.000      42.490  8,475.000       121,501.000     83.260  2,700.000       1.850   339.000 Clusters of cases
The United Kingdom      Europe  4,484,060.000   6,605.280       21,518.000      31.700  3,111.000       127,781.000     188.230 60.000  0.090   6.000   Community transmission
Italy   Europe  4,216,003.000   7,068.910       23,820.000      39.940  2,948.000       126,046.000     211.340 821.000 1.380   44.000  Clusters of cases
Argentina       Americas        3,753,609.000   8,305.220       214,125.000     473.770 21,346.000      77,456.000      171.380 3,393.000       7.510   348.000 Community transmission
Germany Europe  3,681,126.000   4,426.200       29,486.000      35.450  1,978.000       88,442.000      106.340 1,019.000       1.230   36.000  Community transmission
Spain   Europe  3,663,176.000   7,739.220       13,502.000      28.530  0.000   79,888.000      168.780 54.000  0.110   0.000   Community transmission
Colombia        Americas        3,383,279.000   6,649.150       150,823.000     296.410 20,218.000      88,282.000      173.500 3,558.000       6.990   535.000 Community transmission
Iran (Islamic Republic of)      Eastern Mediterranean   2,913,136.000   3,468.310       69,613.000      82.880  11,042.000      80,156.000      95.430  1,308.000       1.560   217.000 Community transmission
Poland  Europe  2,872,283.000   7,566.980       6,102.000       16.080  333.000 73,745.000      194.280 800.000 2.110   7.000   Community transmission
Mexico  Americas        2,412,810.000   1,871.370       16,206.000      12.570  1,307.000       223,507.000     173.350 1,860.0001.440   52.000  Community transmission
Ukraine Europe  2,202,494.000   5,036.140       18,639.000      42.620  1,022.000       50,536.000      115.550 1,100.000       2.520   64.000  Community transmission
Peru    Americas        1,955,469.000   5,930.720       30,180.000      91.530  3,818.000       69,342.000      210.310 1,289.0003.910   140.000 Community transmission
Indonesia       South-East Asia 1,821,703.000   666.010 40,576.000      14.830  5,662.000       50,578.000      18.490  1,123.0000.410   174.000 Community transmission
South Africa    Africa  1,662,825.000   2,803.680       27,360.000      46.130  3,755.000       56,439.000      95.160  637.000 1.070   76.000  Community transmission
Czechia Europe  1,661,272.000   15,534.710      3,180.000       29.740  113.000 30,108.000      281.540 80.000  0.750   4.000   Community transmission
Netherlands     Europe  1,647,418.000   9,463.790       21,424.000      123.070 2,785.000       17,621.000      101.230 79.000  0.450   6.000   Community transmission
Chile   Americas        1,384,346.000   7,241.740       49,085.000      256.770 6,839.000       29,300.000      153.270 752.000 3.930   132.000 Community transmission
Canada  Americas        1,378,971.000   3,653.660       19,112.000      50.640  2,237.000       25,512.000      67.600  281.000 0.740   34.000  Community transmission
```

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, SparkConf

conf = SparkConf().setAppName("ASEAN Case Count")
sc = SparkContext(conf=conf)

input_data = sc.textFile("hdfs://localhost:9000/user/hohuuan/lab04/input/")

filtered_data = input_data.map(lambda line: line.split("\t")) \
.filter(lambda columns: len(columns) >= 3) \
.filter(lambda columns: columns[1] == 'South-East Asia')

total_cases = filtered_data.map(lambda columns: int(columns[2].replace(',', '') \
.replace('.', '')))
total_cases_sum = total_cases.sum()

print("Total Cumulative Cases among ASEAN Countries:", total_cases_sum)

sc.stop()
```

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

```
hohuuan@ubuntu:~/Desktop/spark-3.1.1-bin-hadoop3.2$ spark-submit /home/hohuuan/Desktop/ASEANCaseCount.py > /home/hohuuan/Desktop/output.txt
24/01/28 01:11:17 WARN Utils: Your hostname, ubuntu resolves to a loopback address: 127.0.1.1; using 10.0.2.15 instead (on interface enp0s3)
24/01/28 01:11:17 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
24/01/28 01:11:17 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
24/01/28 01:11:17 INFO SparkContext: Running Spark version 3.1.1
24/01/28 01:11:17 INFO ResourceUtils: ==============================================================
24/01/28 01:11:17 INFO ResourceUtils: No custom resources configured for spark.driver.
24/01/28 01:11:17 INFO ResourceUtils: ==============================================================
24/01/28 01:11:17 INFO SparkContext: Submitted application: ASEAN Case Count
24/01/28 01:11:17 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 1, script: , vendor: , memory -> name: memory, amount: 1024, script: , vendor: , off
Heap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
24/01/28 01:11:17 INFO ResourceProfile: Limiting resource is cpu
24/01/28 01:11:17 INFO ResourceProfileManager: Added ResourceProfile id: 0
24/01/28 01:11:17 INFO SecurityManager: Changing view acls to: hohuuan
24/01/28 01:11:17 INFO SecurityManager: Changing modify acls to: hohuuan
24/01/28 01:11:17 INFO SecurityManager: Changing view acls groups to:
24/01/28 01:11:17 INFO SecurityManager: Changing modify acls groups to:
24/01/28 01:11:17 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(hohuuan); groups with view permissions: Set(); users  with modify permissions
: Set(hohuuan); groups with modify permissions: Set()
24/01/28 01:11:18 INFO Utils: Successfully started service 'sparkDriver' on port 43341.
24/01/28 01:11:18 INFO SparkEnv: Registering MapOutputTracker
24/01/28 01:11:18 INFO SparkEnv: Registering BlockManagerMaster
24/01/28 01:11:18 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
24/01/28 01:11:18 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
24/01/28 01:11:18 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
24/01/28 01:11:18 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-11598f93-b18e-44d4-8e20-8e48b7876ead
24/01/28 01:11:18 INFO MemoryStore: MemoryStore started with capacity 366.3 MiB
24/01/28 01:11:18 INFO SparkEnv: Registering OutputCommitCoordinator
24/01/28 01:11:18 INFO Utils: Successfully started service 'SparkUI' on port 4040.
24/01/28 01:11:18 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://10.0.2.15:4040
24/01/28 01:11:18 INFO Executor: Starting executor ID driver on host 10.0.2.15
24/01/28 01:11:18 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 44973.
24/01/28 01:11:18 INFO NettyBlockTransferService: Server created on 10.0.2.15:44973
24/01/28 01:11:18 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
24/01/28 01:11:18 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, 10.0.2.15, 44973, None)
24/01/28 01:11:18 INFO BlockManagerMasterEndpoint: Registering block manager 10.0.2.15:44973 with 366.3 MiB RAM, BlockManagerId(driver, 10.0.2.15, 44973, None)
24/01/28 01:11:18 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, 10.0.2.15, 44973, None)
24/01/28 01:11:18 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, 10.0.2.15, 44973, None)
24/01/28 01:11:19 INFO MemoryStore: Block broadcast_0 stored as values in memory (estimated size 293.9 KiB, free 366.0 MiB)
24/01/28 01:11:19 INFO MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 27.0 KiB, free 366.0 MiB)
24/01/28 01:11:19 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on 10.0.2.15:44973 (size: 27.0 KiB, free: 366.3 MiB)
24/01/28 01:11:19 INFO SparkContext: Created broadcast 0 from textFile at NativeMethodAccessorImpl.java:0
```

```
hohuuan@ubuntu:~/Desktop/spark-3.1.1-bin-hadoop3.2$ cat /home/hohuuan/Desktop/output.txt
Total Cumulative Cases among ASEAN Countries: 31923614000
```

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).