Course: Big Data

Lab 03

MapReduce

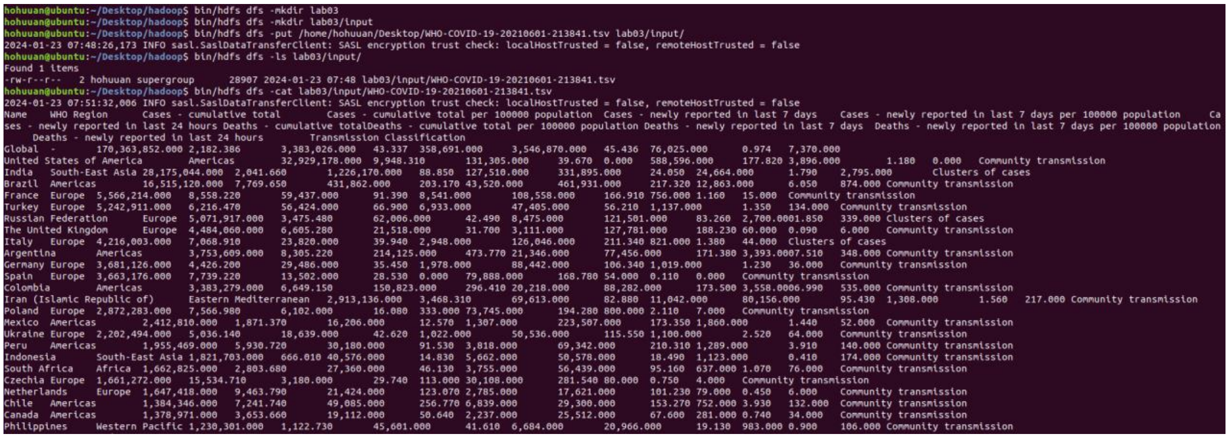
Fill answers of the questions below in the given tables.
Your screenshots must contain commands for required operations.

Question 1:

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 **lab03**, in HDFS and then copy the tsv to **lab03/input/**

Take a screenshot to show the content of **lab03/input/** in HDFS



```
hohuuan@ubuntu:~/Desktop/hadoop$ bin/hdfs dfs -mkdir lab03
hohuuan@ubuntu:~/Desktop/hadoop$ bin/hdfs dfs -mkdir lab03/input
hohuuan@ubuntu:~/Desktop/hadoop$ bin/hdfs dfs -put /home/hohuuan/Desktop/WHO-COVID-19-20210601-213841.tsv lab03/input/
2024-01-23 07:48:26,173 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
hohuuan@ubuntu:~/Desktop/hadoop$ bin/hdfs dfs -ls lab03/input/
Found 1 items
-rw-r--r--   2 hohuuan supergroup      28907 2024-01-23 07:48 lab03/input/WHO-COVID-19-20210601-213841.tsv
hohuuan@ubuntu:~/Desktop/hadoop$ bin/hdfs dfs -cat lab03/input/WHO-COVID-19-20210601-213841.tsv
2024-01-23 07:51:32,006 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 totalDeaths - cumulative total per 100000 population Deaths - newly reported in last 7 days  Deaths - newly reported in last 7 days per 100000 population
        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  2,700.0001.850  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 3,393.0007.510  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 3,558.0006.990  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 1,860.000       1.440   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 1,289.000       3.910   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  1,123.000       0.410   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
Philippines     Western Pacific 1,230,301.000   1,122.730       45,601.000      41.610  6,684.000       20,966.000      19.130  983.000 0.900   106.000 Community transmission
```

Question 2:

Create one and only one java file, named **ASEANCaseCount.java**, to run a MapReduce job that counts the number of cumulative total cases among ASEAN countries (*South-East Asia Region in the given data table*).
The output of the MapReduce job is located in **lab03/output-java/**.
Submit the source code file following the instructions in Submission Notice.

Question 3 (*optional*):

Create a pair of Python files, named **ASEANDeathCountMapper.py** and **ASEANDeathCountReducer.py**, to run a MapReduce job that counts the number of cumulative total deaths among ASEAN countries (*South-East Asia Region in the given data table*).
The output of the MapReduce job is located in **lab03/output-python/**.
Submit the source code files following the instructions in Submission Notice.

Submission Notice

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

  lab03_<student number>_HoTen.pdf

  E.g. lab03_123456_NguyenThanhAn.pdf

  If you have not been assigned a student number yet, then use 123456 instead.
- Create a folder with the name as **<student number>_HoTen**, which contains
  - **<student number>_HoTen.pdf** → your answer
  - **java/** **|** → Java source code folder
    | ASEANCaseCount.java
  - **python/** | → Python source code folder
    | **ASEANDeathCountMapper.py**
    | ASEANDeathCountReducer.py
- Compress the folder **<student number>_HoTen** in zip format and finally submit to the given form.

  E.g. 123456_HoTen.zip
- Careless mistakes in filename, format, question order, etc. are not accepted (0 pts).