ASSIGNMENT 2 - MAPREDUCE

Affan Mohammed N Marikar
281911

Task 1

mapper1.py

```
#!/usr/bin/env python

import sys
import csv

def read_input(file):
        for row in csv.reader(file):
        yield row

# Read input
data = read_input(sys.stdin)

# Skip the header row
header = next(data)

for row in data:
   zone = row[4]  # 'zone'
   regional_zone = row[5]  # 'WH_regional_zone'
   product_wg_ton = float(row[22])  # 'product_wg_ton'

        # Emit key-value pair
   print(f"{zone}\t{regional_zone}\t{product_wg_ton}")
```

reducer1.py

```
#!/usr/bin/env python
import sys

def read_input(file):
        for line in file:
        yield line.strip().split('\t')

current_zone = None
current_regional_zone = None
current_total_weight = 0

data = read_input(sys.stdin)

for line in data:
  if len(line) != 3:
```

```
            continue  # Skip malformed lines

    zone, regional_zone, weight = line

    try:
            weight = float(weight)
    except ValueError:
            continue  # Skip lines with invalid data

    if (current_zone == zone) and (current_regional_zone == regional_zone):
            current_total_weight += weight
    else:
            if current_zone and current_regional_zone:
            # Emit the total weight for the previous zone and regional zone
            print(f"{current_zone}\t{current_regional_zone}\t{current_total_weight}")

            current_zone = zone
            current_regional_zone = regional_zone
            current_total_weight = weight

# Emit the total weight for the last zone and regional zone
if current_zone and current_regional_zone:
            print(f"{current_zone}\t{current_regional_zone}\t{current_total_weight}")
```

Output

cat FMCG_data.csv | python3 mapper1.py | sort -k1,1 | python3 reducer1.py

```
East    Zone 1  1153.0
East    Zone 3  3024.0
East    Zone 4  3692.0
East    Zone 5  1036.0
East    Zone 6  870.0
North   Zone 1  16191.0
North   Zone 2  10027.0
North   Zone 3  24339.0
North   Zone 4  30836.0
North   Zone 5  35000.0
North   Zone 6  88434.0
South   Zone 1  14325.0
South   Zone 2  35969.0
South   Zone 3  18983.0
South   Zone 4  20364.0
South   Zone 5  19027.0
South   Zone 6  20276.0
West    Zone 1  9832.0
West    Zone 2  6702.0
West    Zone 3  18609.0
West    Zone 4  35636.0
West    Zone 5  20804.0
West    Zone 6  35178.0
```

Task 2

mapper2.py

```
#!/usr/bin/python3
"""mapper.py"""

import sys
import csv

for row in csv.reader(sys.stdin):
        print("%s\t%s"%(row[3],row[6]))
```

reducer2.py

```
#!/usr/bin/python3
"""reducer.py"""

import sys
```

```
#!/usr/bin/python3
"""reducer.py"""
import sys
import pandas as pd
wh={'capacity':[], 'refill':[]}
for line in sys.stdin:
 capacity, refill = line.strip().split("\t")
 try:
   refill = int(refill)
 except ValueError:
   continue
 wh['capacity'].append(capacity)
 wh['refill'].append(refill)
df = pd.DataFrame(wh)
df['int_capacity']=pd.factorize(df['capacity'])[0]
correla =df[['int_capacity','refill']].corr()
print(correla)
```

Output

cat FMCG_data.csv | python3 mapper2.py | sort -k1,1 | python3 reducer2b.py

```
hadoop@hadoop-VirtualBox:~/Assi2$ cat FMCG_data.csv | python3 mapper2.py | sort
-k1,1 | python3 reducer2b.py
              int_capacity    refill
int_capacity      1.000000 -0.007054
refill           -0.007054  1.000000
hadoop@hadoop-VirtualBox:~/Assi2$
```

Task3

mapper3.py

```
#!/usr/bin/python3
"""mapper.py"""

import sys
import csv

for row in csv.reader(sys.stdin):
        print("%s\t%s"%(row[7],row[23]))
```

reducer3.py

```
#!/usr/bin/python3
"""reducer.py"""

import sys

dict={}

for line in sys.stdin:
        trans, weight = line.strip().split("\t")
        try:
        weight = float(weight)
        except ValueError:
        continue
        if trans in dict:
                dict[trans]+=weight
        else:
        dict[trans]=weight

print("transport\tweight")
for i in dict:
        print("%s\t\t%s"%(i, dict[i]))
```

Output

```
cat FMCG_data.csv | python3 mapper3.py | sort -k1,1 | python3 reducer3.py
```

```
transport       weight
0               359167349.0
1               99133868.0
2               41450553.0
3               32129593.0
4               14896451.0
5               5788009.0
```

Task 4

mapper.py

```
#!/usr/bin/python3
"""mapper.py"""

import sys
import csv

for row in csv.reader(sys.stdin):
        print("%s\t%s"%(row[18],row[23]))
```

reducer.py

```
#!/usr/bin/python3
"""reducer.py"""

import sys

dict={}

for line in sys.stdin:
        storage, weight = line.strip().split("\t")
        try:
        weight = float(weight)
        except ValueError:
        continue
        if storage in dict:
                dict[storage].append(weight)
        else:
        dict[storage]=[weight]

print("issue\tweight\t\tagg")
for i in dict:
        print("%s\t%s\t%s"%(i, sum(dict[i]), sum(dict[i])/len(dict[i]) ))
```

Output

```
cat FMCG_data.csv | python3 mapper4.py | sort -k1,1 | python3 reducer4.py
```

```
hadoop@hadoop-VirtualBox:~/Assi2$ cat FMCG_data.cs
issue    weight           agg
0        4930869.0        5430.472466960352
10       8259859.0        12966.811616954474
11       12270859.0       14153.239907727797
12       11436927.0       15476.220568335588
13       12163798.0       16754.54269972452
14       14535116.0       17704.16077953715
15       17281171.0       19032.12665198238
16       19200310.0       20469.413646055436
17       16416984.0       21918.536715620827
18       24289887.0       22700.828971962615
19       24569176.0       24040.28962818004
20       27006058.0       25357.800938967135
21       18581712.0       27047.615720524016
22       25472459.0       27930.327850877195
23       26797528.0       29223.040348964012
24       42904667.0       30129.681882022473
25       39461458.0       31268.984152139463
26       19958755.0       32772.99671592775
27       19849883.0       33931.42393162393
28       12281089.0       36550.86011904762
29       12068423.0       37596.333333333336
30       13109614.0       38900.93175074184
31       11698085.0       40477.80276816609
32       12244881.0       41367.84121621621
33       12650336.0       42882.49491525424
34       12750651.0       44273.09375
```