Big Data Processing Lab-1

Roll no. 202201320

202201320_colab_link:

https://colab.research.google.com/drive/1nCT2L7657xI_FWtTtF5Oq_hMrqmqIzzZ?usp=sharing

Exercise 1: Go through the map-reduce programs in the given google colab notebook.

1) Word count:

Mapper: It takes a string as an input and the mapper yields the unique words with a one value indicating that that word is present once.

Reducer: Word and count from the string are the key value inputs to the reducer. The reducer returns the word and the sum(count) that is the total number of times the particular word occurs in the given input file.

2) Salary grouped by Department_no:

Mapper: The records of the csv file is the input to the mapper. It returns only the Department_no and the salary of each record as key value pair.

Reducer: The reducer takes the Department_no and salary as input and then output the sum of salary which is sum(salary) and returns it as key value pair, where key is Department_no and value is the sum of all employees' salary in that particular department.

3) Image counter from web log Mapper:

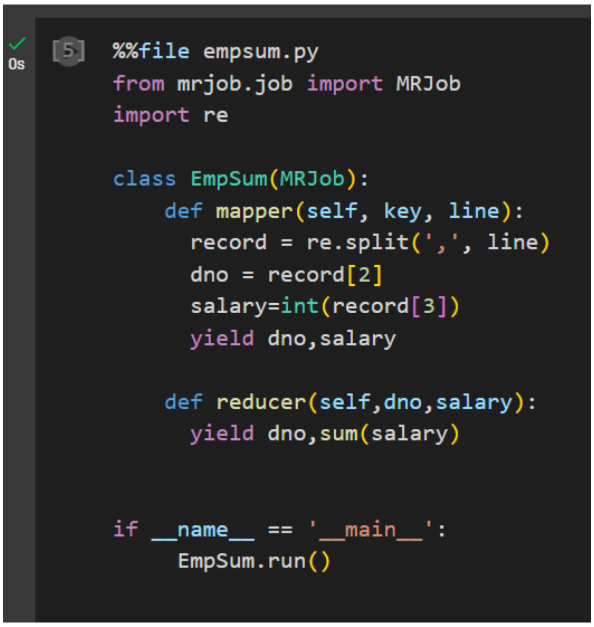
Mapper: Each line is an input to the mapper. The line is split into multiple categories and the mapper check each tag and segregates it. Then, it returns the category and count as the key value in string integer format.

Reducer: It gets the category and counts from the mapper and counts the input and return each category and its count in the key value pair.

Exercise 2: Process “employee.csv” Write down map-reduce programs for performing operations that are equivalent to following SQL statements on given employee data file “employee.csv”.

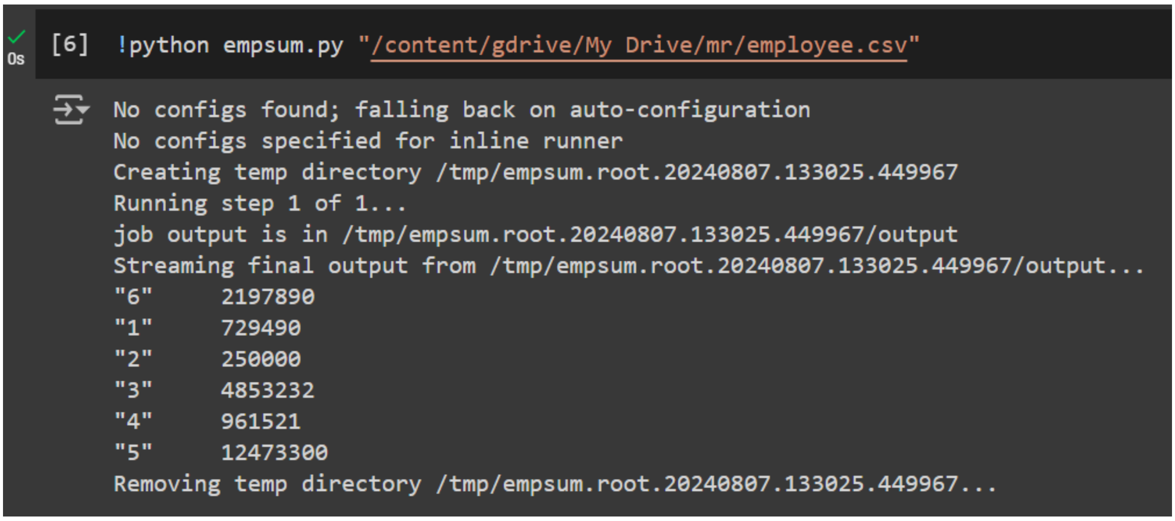
1. Compute department wise total salary.

Code:



```
%%file empsum.py
from mrjob.job import MRJob
import re

class EmpSum(MRJob):
    def mapper(self, key, line):
      record = re.split(',', line)
      dno = record[2]
      salary=int(record[3])
      yield dno,salary

    def reducer(self,dno,salary):
      yield dno,sum(salary)


if __name__ == '__main__':
     EmpSum.run()
```

Output:

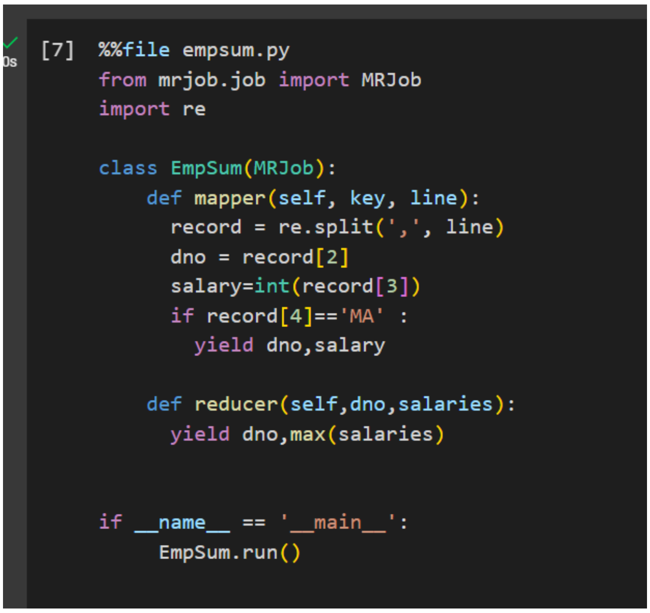


```
!python empsum.py "/content/gdrive/My Drive/mr/employee.csv"
```

```
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/empsum.root.20240807.133025.449967
Running step 1 of 1...
job output is in /tmp/empsum.root.20240807.133025.449967/output
Streaming final output from /tmp/empsum.root.20240807.133025.449967/output...
"6"     2197890
"1"     729490
"2"     250000
"3"     4853232
"4"     961521
"5"     12473300
Removing temp directory /tmp/empsum.root.20240807.133025.449967...
```

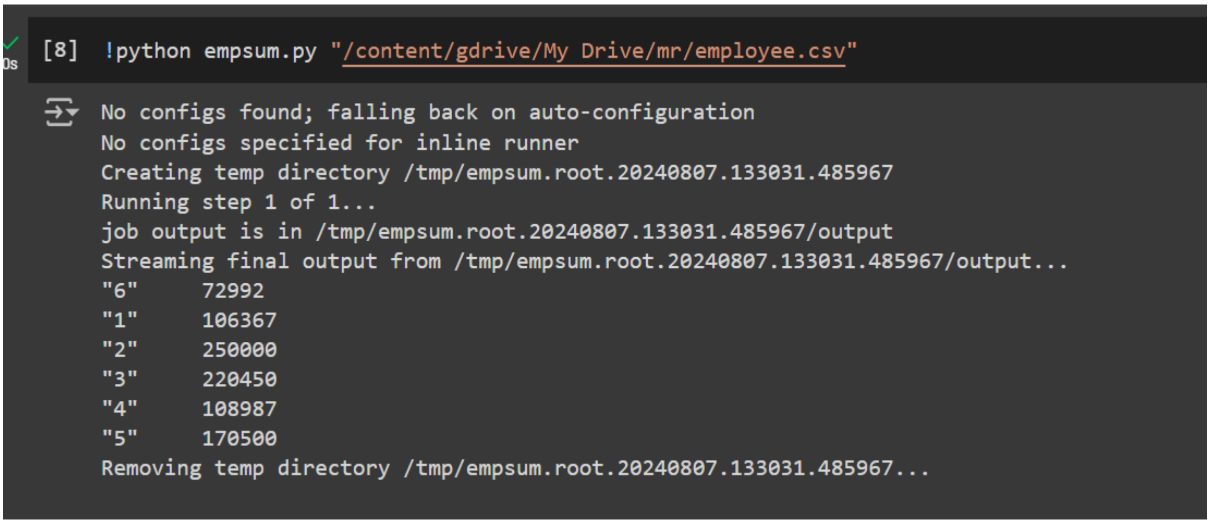
2.Compute department wise maximum salary among employees from Massachusetts (’MA’).

Code:



```
%%file empsum.py
from mrjob.job import MRJob
import re

class EmpSum(MRJob):
    def mapper(self, key, line):
      record = re.split(',', line)
      dno = record[2]
      salary=int(record[3])
      if record[4]=='MA' :
        yield dno,salary

    def reducer(self,dno,salaries):
      yield dno,max(salaries)


if __name__ == '__main__':
     EmpSum.run()
```

Output:

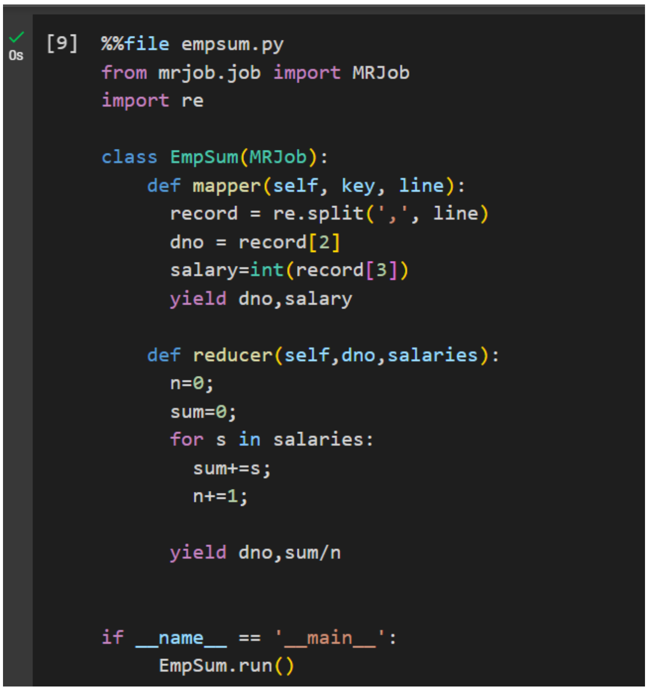


```
!python empsum.py "/content/gdrive/My Drive/mr/employee.csv"
```

```
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/empsum.root.20240807.133031.485967
Running step 1 of 1...
job output is in /tmp/empsum.root.20240807.133031.485967/output
Streaming final output from /tmp/empsum.root.20240807.133031.485967/output...
"6"     72992
"1"     106367
"2"     250000
"3"     220450
"4"     108987
"5"     170500
Removing temp directory /tmp/empsum.root.20240807.133031.485967...
```

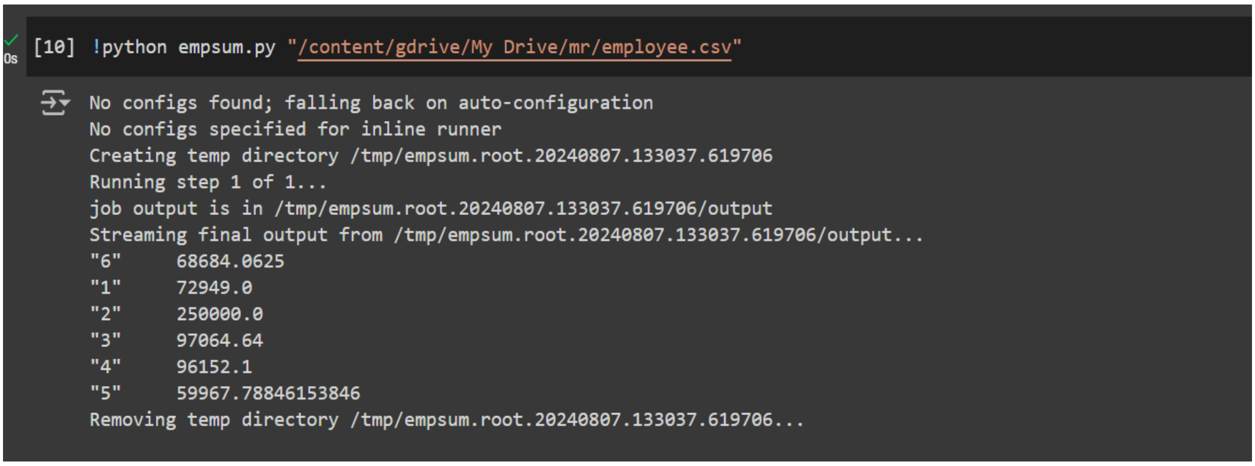
3.Compute department wise average salary.

Code:



```
%%file empsum.py
from mrjob.job import MRJob
import re

class EmpSum(MRJob):
    def mapper(self, key, line):
      record = re.split(',', line)
      dno = record[2]
      salary=int(record[3])
      yield dno,salary

    def reducer(self,dno,salaries):
      n=0;
      sum=0;
      for s in salaries:
        sum+=s;
        n+=1;

      yield dno,sum/n


if __name__ == '__main__':
     EmpSum.run()
```

Output:

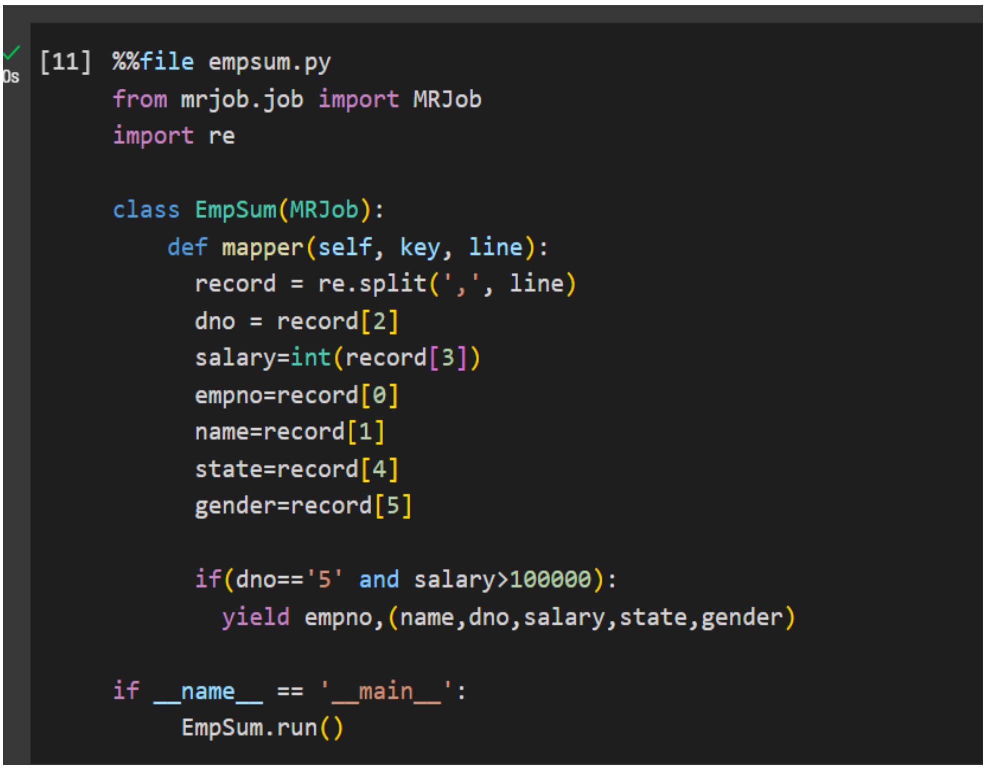


```
!python empsum.py "/content/gdrive/My Drive/mr/employee.csv"
```

```
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/empsum.root.20240807.133037.619706
Running step 1 of 1...
job output is in /tmp/empsum.root.20240807.133037.619706/output
Streaming final output from /tmp/empsum.root.20240807.133037.619706/output...
"6"     68684.0625
"1"     72949.0
"2"     250000.0
"3"     97064.64
"4"     96152.1
"5"     59967.78846153846
Removing temp directory /tmp/empsum.root.20240807.133037.619706...
```

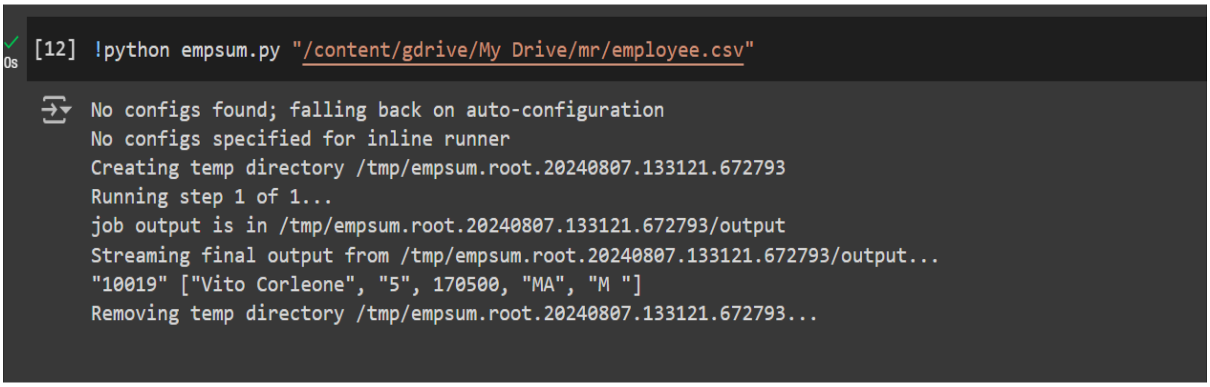
4. List all details of employees that are from dno=5 and salary is greater than 100 thousand.

Code:



```
%%file empsum.py
from mrjob.job import MRJob
import re

class EmpSum(MRJob):
    def mapper(self, key, line):
      record = re.split(',', line)
      dno = record[2]
      salary=int(record[3])
      empno=record[0]
      name=record[1]
      state=record[4]
      gender=record[5]

      if(dno=='5' and salary>100000):
        yield empno,(name,dno,salary,state,gender)

if __name__ == '__main__':
     EmpSum.run()
```

Output:

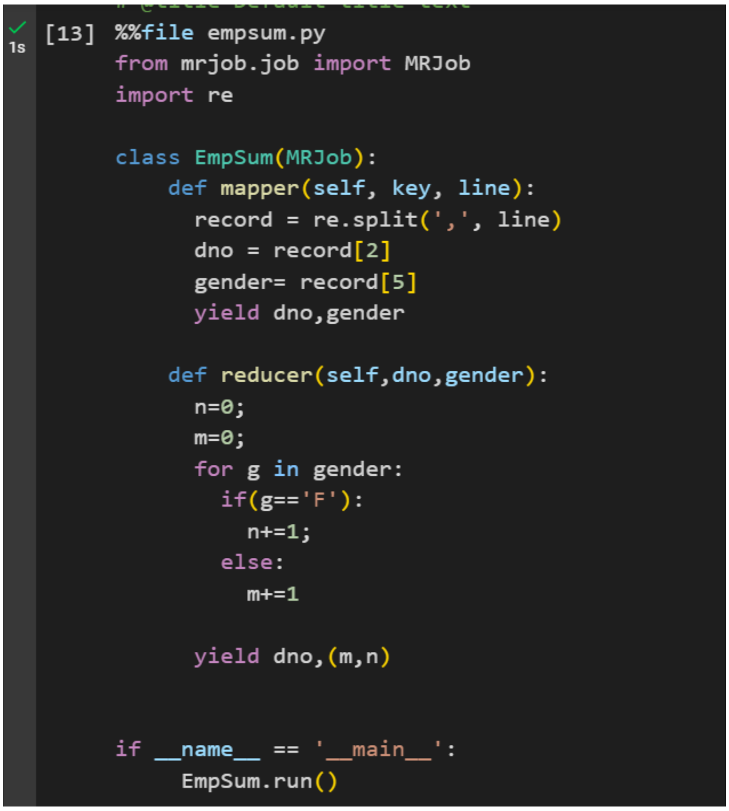


```
!python empsum.py "/content/gdrive/My Drive/mr/employee.csv"
```

```
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/empsum.root.20240807.133121.672793
Running step 1 of 1...
job output is in /tmp/empsum.root.20240807.133121.672793/output
Streaming final output from /tmp/empsum.root.20240807.133121.672793/output...
"10019" ["Vito Corleone", "5", 170500, "MA", "M "]
Removing temp directory /tmp/empsum.root.20240807.133121.672793...
```

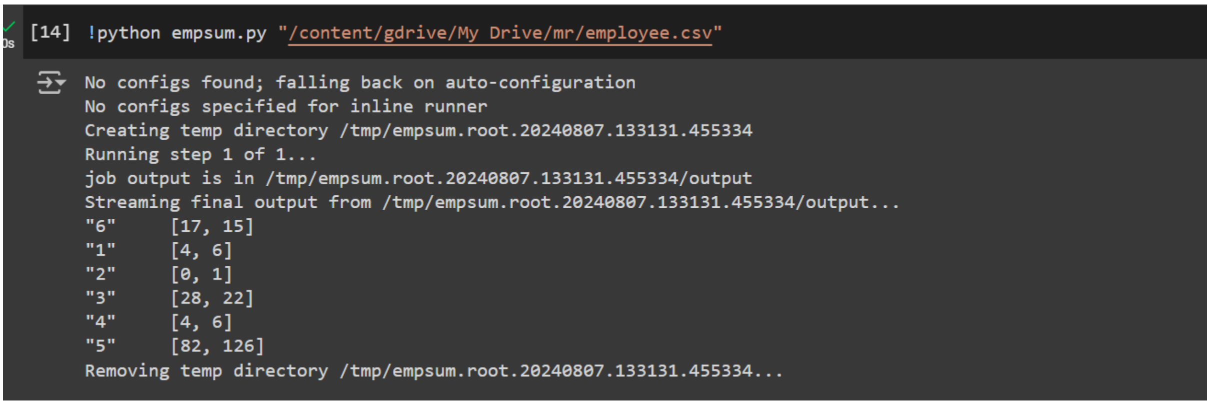
5. Compute total number of male and female employees for each department.

Code:



```
%%file empsum.py
from mrjob.job import MRJob
import re

class EmpSum(MRJob):
    def mapper(self, key, line):
      record = re.split(',', line)
      dno = record[2]
      gender= record[5]
      yield dno,gender

    def reducer(self,dno,gender):
      n=0;
      m=0;
      for g in gender:
        if(g=='F'):
          n+=1;
        else:
          m+=1

      yield dno,(m,n)


if __name__ == '__main__':
     EmpSum.run()
```

Output:

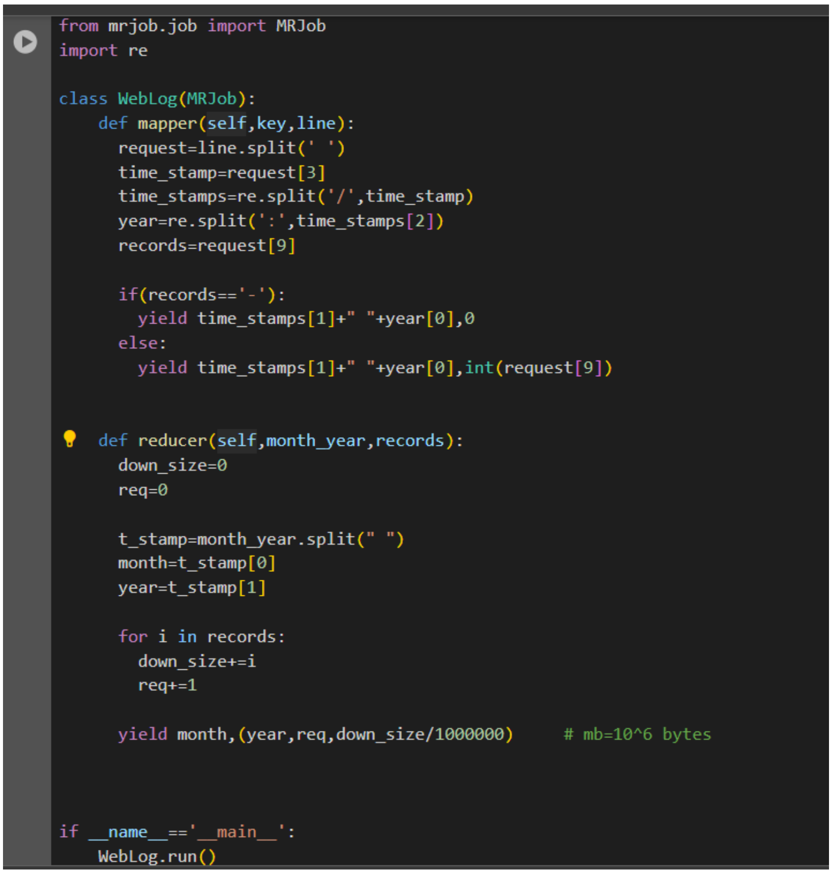


```
!python empsum.py "/content/gdrive/My Drive/mr/employee.csv"

No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/empsum.root.20240807.133131.455334
Running step 1 of 1...
job output is in /tmp/empsum.root.20240807.133131.455334/output
Streaming final output from /tmp/empsum.root.20240807.133131.455334/output...
"6"     [17, 15]
"1"     [4, 6]
"2"     [0, 1]
"3"     [28, 22]
"4"     [4, 6]
"5"     [82, 126]
Removing temp directory /tmp/empsum.root.20240807.133131.455334...
```

Exercise 2: Process "web access log" A web server produces a log file that has entries for every http request it receives to a resource.

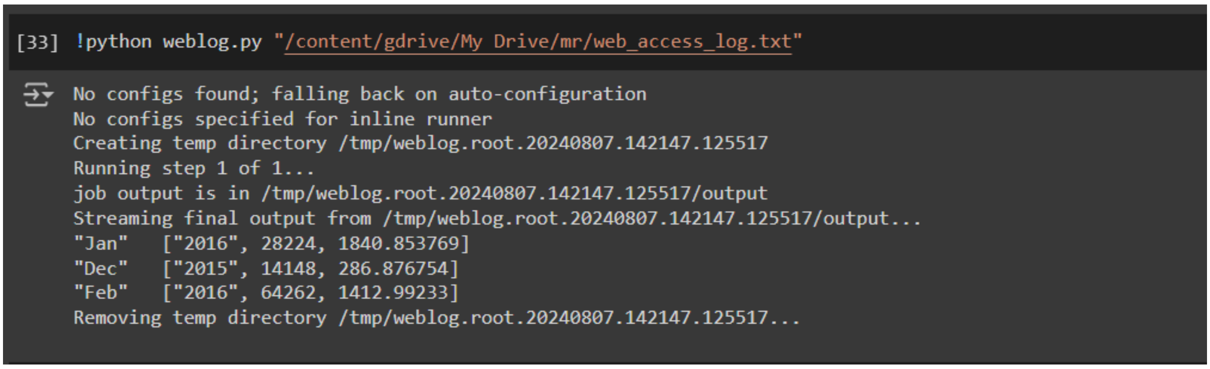
6. Computes following summaries on a Monthly Basis –
 (a) Total number of requests.
 (b) Total download size (in Megabytes)

Code:



```
from mrjob.job import MRJob
import re

class WebLog(MRJob):
    def mapper(self,key,line):
      request=line.split(' ')
      time_stamp=request[3]
      time_stamps=re.split('/',time_stamp)
      year=re.split(':',time_stamps[2])
      records=request[9]

      if(records=='-'):
        yield time_stamps[1]+" "+year[0],0
      else:
        yield time_stamps[1]+" "+year[0],int(request[9])


    def reducer(self,month_year,records):
      down_size=0
      req=0

      t_stamp=month_year.split(" ")
      month=t_stamp[0]
      year=t_stamp[1]

      for i in records:
        down_size+=i
        req+=1

      yield month,(year,req,down_size/1000000)      # mb=10^6 bytes



if __name__=='__main__':
    WebLog.run()
```

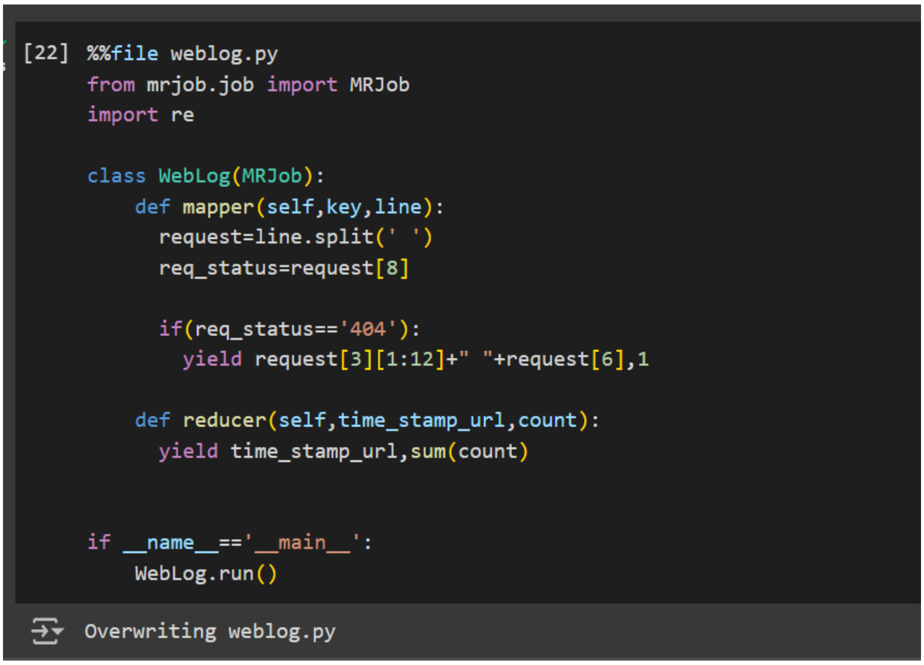
Output:



```
[33] !python weblog.py "/content/gdrive/My Drive/mr/web_access_log.txt"

No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/weblog.root.20240807.142147.125517
Running step 1 of 1...
job output is in /tmp/weblog.root.20240807.142147.125517/output
Streaming final output from /tmp/weblog.root.20240807.142147.125517/output...
"Jan"   ["2016", 28224, 1840.853769]
"Dec"   ["2015", 14148, 286.876754]
"Feb"   ["2016", 64262, 1412.99233]
Removing temp directory /tmp/weblog.root.20240807.142147.125517...
```

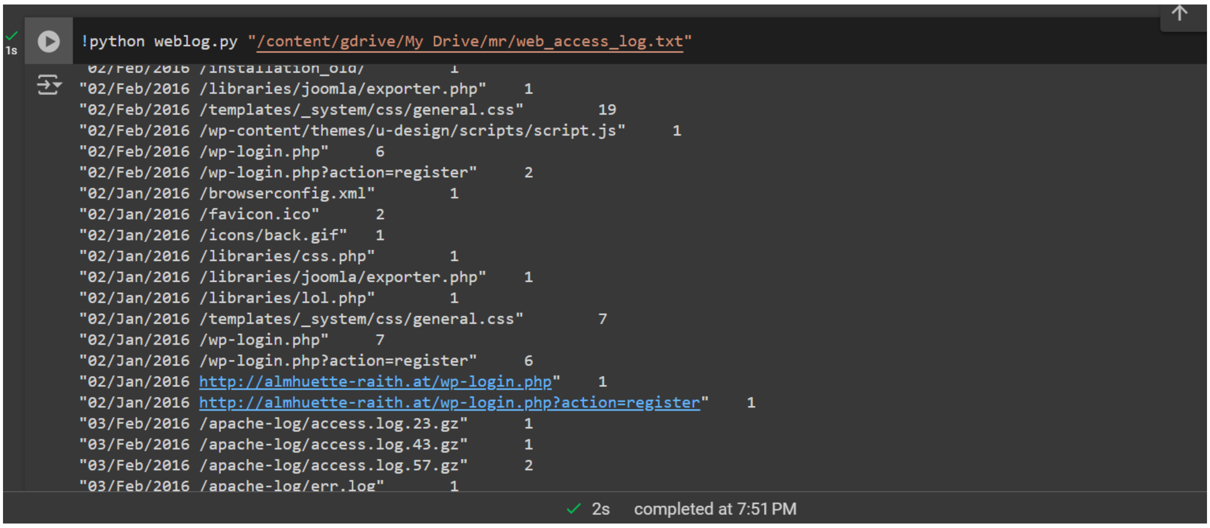
7. Create another Map Reduce program that lists Timestamp, and URL for which http response status has been 404.

Code:



```
[22] %%file weblog.py
     from mrjob.job import MRJob
     import re

     class WebLog(MRJob):
         def mapper(self,key,line):
           request=line.split(' ')
           req_status=request[8]

           if(req_status=='404'):
             yield request[3][1:12]+" "+request[6],1

         def reducer(self,time_stamp_url,count):
           yield time_stamp_url,sum(count)


     if __name__=='__main__':
         WebLog.run()

Overwriting weblog.py
```

Output:



```
!python weblog.py "/content/gdrive/My Drive/mr/web_access_log.txt"
"02/Feb/2016 /installation_old/"         1
"02/Feb/2016 /libraries/joomla/exporter.php"    1
"02/Feb/2016 /templates/_system/css/general.css"        19
"02/Feb/2016 /wp-content/themes/u-design/scripts/script.js"     1
"02/Feb/2016 /wp-login.php"     6
"02/Feb/2016 /wp-login.php?action=register"     2
"02/Jan/2016 /browserconfig.xml"        1
"02/Jan/2016 /favicon.ico"      2
"02/Jan/2016 /icons/back.gif"   1
"02/Jan/2016 /libraries/css.php"        1
"02/Jan/2016 /libraries/joomla/exporter.php"    1
"02/Jan/2016 /libraries/lol.php"        1
"02/Jan/2016 /templates/_system/css/general.css"        7
"02/Jan/2016 /wp-login.php"     7
"02/Jan/2016 /wp-login.php?action=register"     6
"02/Jan/2016 http://almhuette-raith.at/wp-login.php"    1
"02/Jan/2016 http://almhuette-raith.at/wp-login.php?action=register"    1
"03/Feb/2016 /apache-log/access.log.23.gz"      1
"03/Feb/2016 /apache-log/access.log.43.gz"      1
"03/Feb/2016 /apache-log/access.log.57.gz"      2
"03/Feb/2016 /apache-log/err.log"       1
```