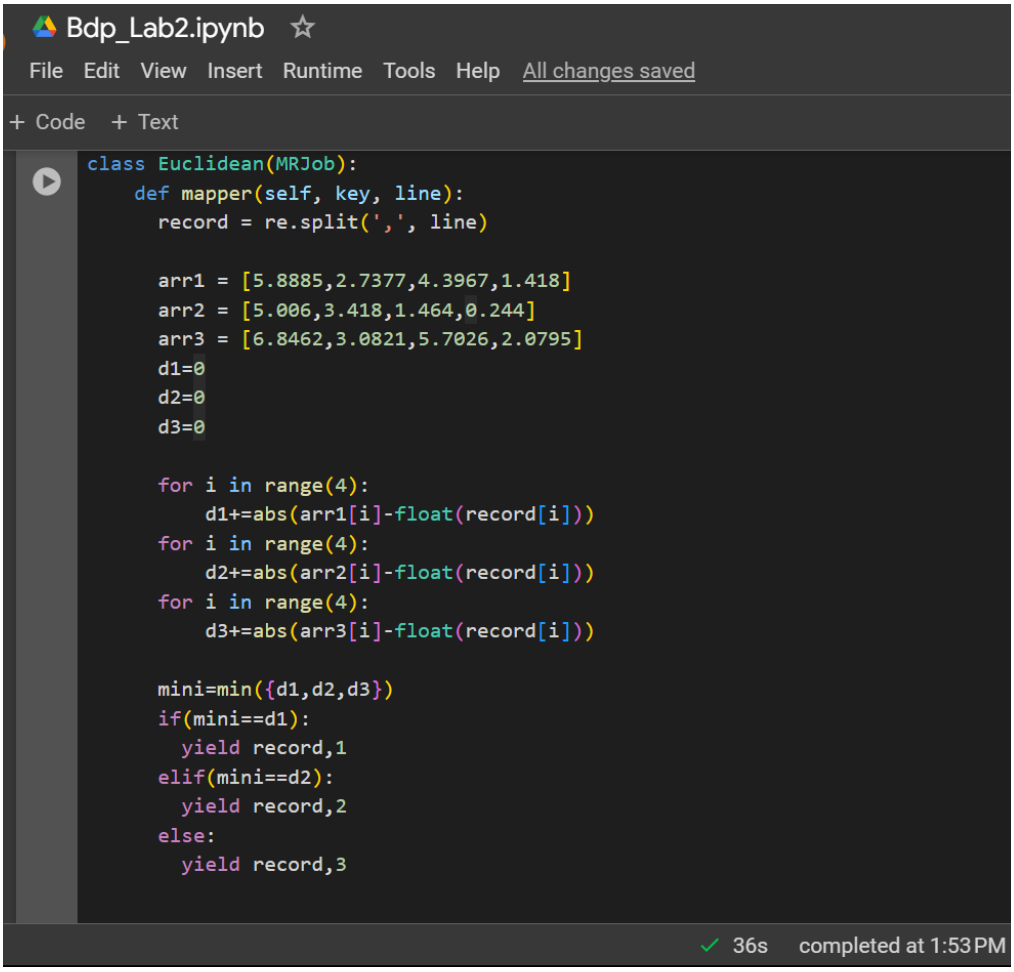
Big Data Processing Lab-2

Roll No – 202201320

Colab Link:
https://colab.research.google.com/drive/1Uosv0KFLZJKmUkzO49KNjZ_1aE7JXGL5?usp=sharing

Exercise #1: Iterate through all data vector and determine nearest class vector for each data vector. Let nearness be computed by Euclidean distance between data vector and class vector. Your program should output ID of class vector for each data vector. Let ID of class vector be 1,2,3 in the order of their occurrence in the class file.

Code:



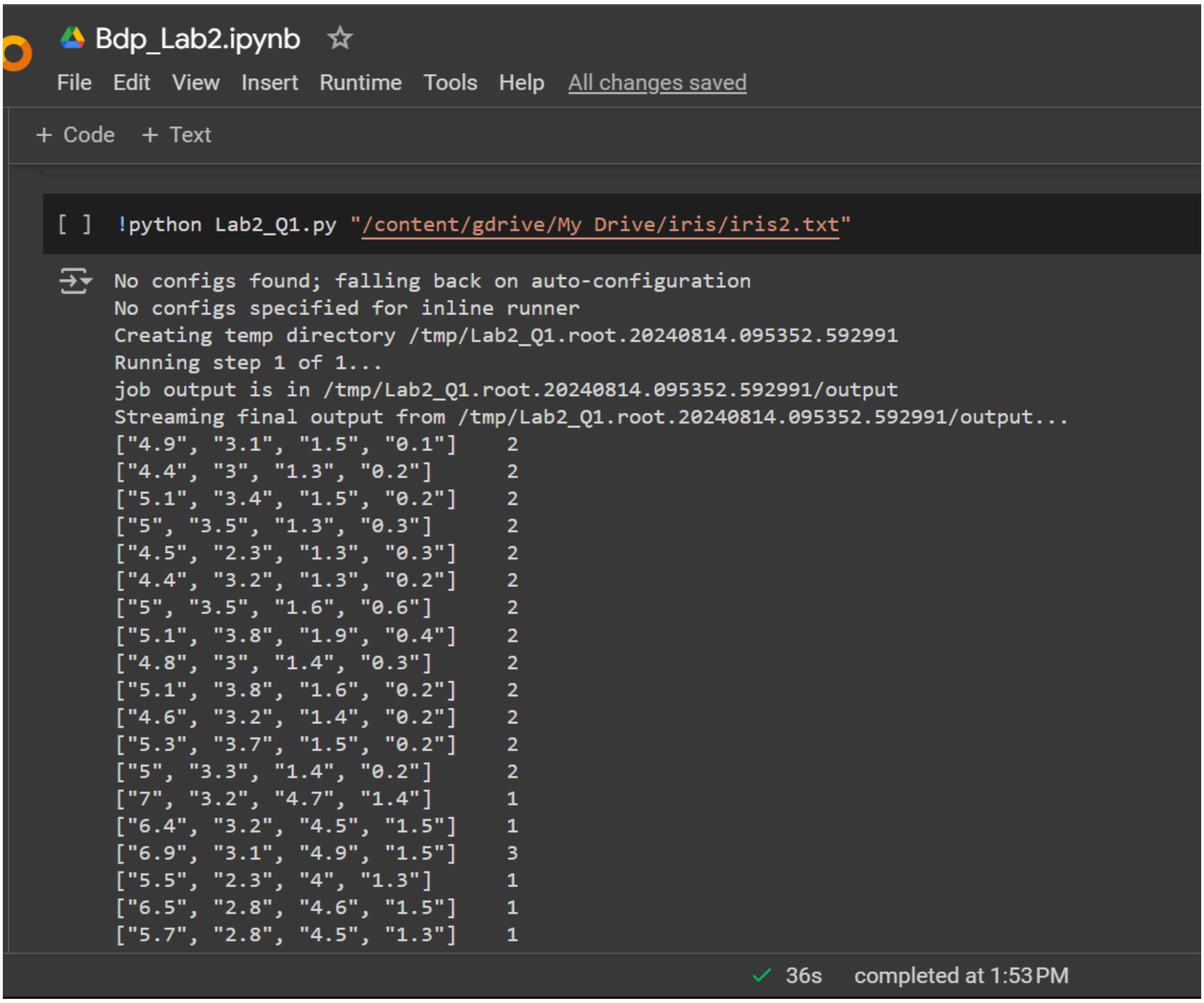
```
class Euclidean(MRJob):
    def mapper(self, key, line):
      record = re.split(',', line)

      arr1 = [5.8885,2.7377,4.3967,1.418]
      arr2 = [5.006,3.418,1.464,0.244]
      arr3 = [6.8462,3.0821,5.7026,2.0795]
      d1=0
      d2=0
      d3=0

      for i in range(4):
          d1+=abs(arr1[i]-float(record[i]))
      for i in range(4):
          d2+=abs(arr2[i]-float(record[i]))
      for i in range(4):
          d3+=abs(arr3[i]-float(record[i]))

      mini=min({d1,d2,d3})
      if(mini==d1):
        yield record,1
      elif(mini==d2):
        yield record,2
      else:
        yield record,3
```

Explanation:

Here, as instructed in the question, the nearness of all the vectors is calculated with the given three vectors and the Euclidean distance is stored in variables d1, d2 and d3.

The minimum of all three is stored in variable mini and is yielded with its corresponding index.

Output:



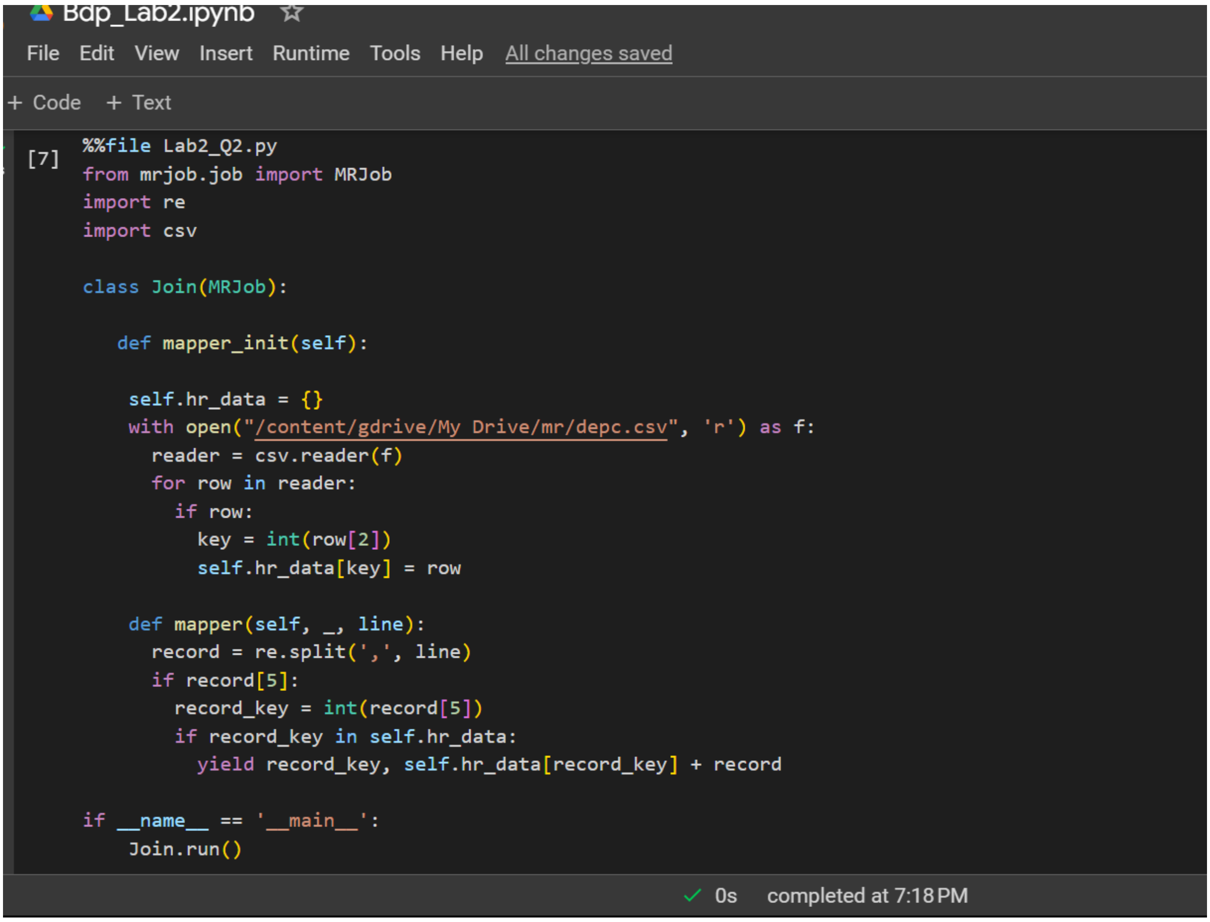
```
[ ] !python Lab2_Q1.py "/content/gdrive/My Drive/iris/iris2.txt"
```

```
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/Lab2_Q1.root.20240814.095352.592991
Running step 1 of 1...
job output is in /tmp/Lab2_Q1.root.20240814.095352.592991/output
Streaming final output from /tmp/Lab2_Q1.root.20240814.095352.592991/output...
["4.9", "3.1", "1.5", "0.1"]    2
["4.4", "3", "1.3", "0.2"]      2
["5.1", "3.4", "1.5", "0.2"]    2
["5", "3.5", "1.3", "0.3"]      2
["4.5", "2.3", "1.3", "0.3"]    2
["4.4", "3.2", "1.3", "0.2"]    2
["5", "3.5", "1.6", "0.6"]      2
["5.1", "3.8", "1.9", "0.4"]    2
["4.8", "3", "1.4", "0.3"]      2
["5.1", "3.8", "1.6", "0.2"]    2
["4.6", "3.2", "1.4", "0.2"]    2
["5.3", "3.7", "1.5", "0.2"]    2
["5", "3.3", "1.4", "0.2"]      2
["7", "3.2", "4.7", "1.4"]      1
["6.4", "3.2", "4.5", "1.5"]    1
["6.9", "3.1", "4.9", "1.5"]    3
["5.5", "2.3", "4", "1.3"]      1
["6.5", "2.8", "4.6", "1.5"]    1
["5.7", "2.8", "4.5", "1.3"]    1
```

Exercise #2: Suppose you are given two files employee "empc.csv" and department "depc.csv", and the attributes of these files are as:

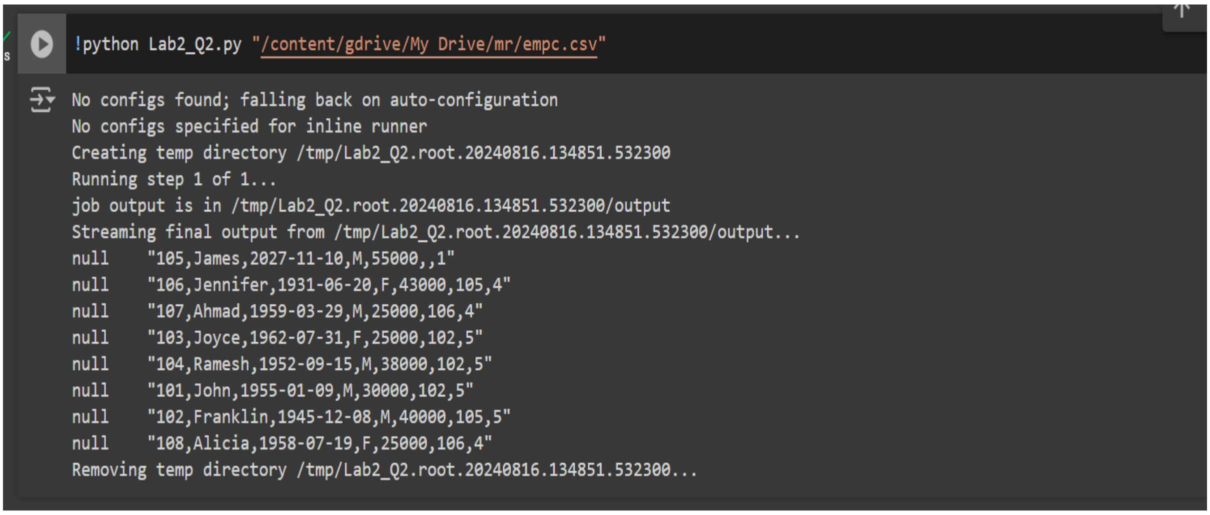
Perform JOIN operation on these two files using the map-reduce approach. Let the joining condition be "mgr_eno=eno"

Code:


Bdp_Lab2.ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
[7]
✓ 0s completed at 7:18 PM

```
%%file Lab2_Q2.py
from mrjob.job import MRJob
import re
import csv

class Join(MRJob):

   def mapper_init(self):

    self.hr_data = {}
    with open("/content/gdrive/My Drive/mr/depc.csv", 'r') as f:
      reader = csv.reader(f)
      for row in reader:
        if row:
          key = int(row[2])
          self.hr_data[key] = row

    def mapper(self, _, line):
      record = re.split(',', line)
      if record[5]:
        record_key = int(record[5])
        if record_key in self.hr_data:
          yield record_key, self.hr_data[record_key] + record

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

Output:



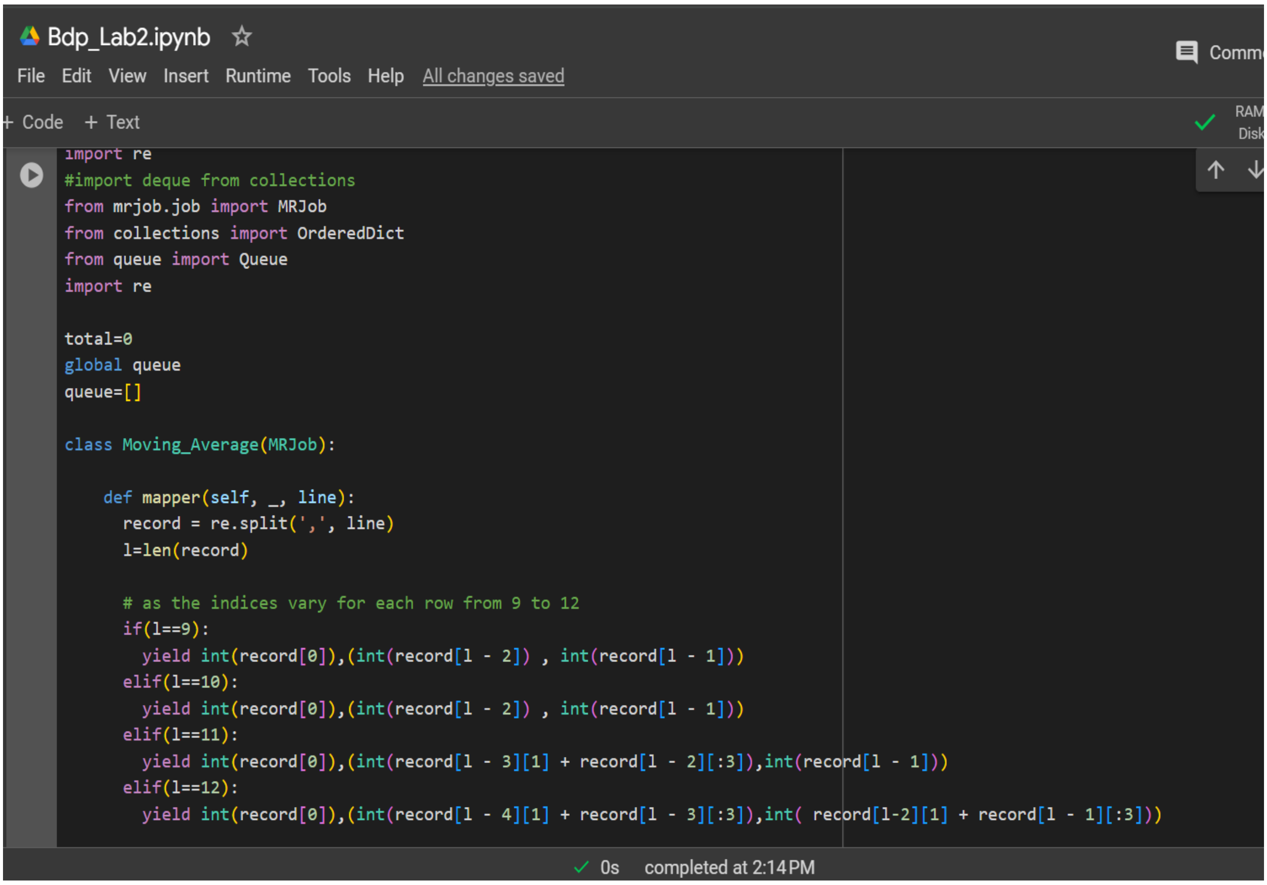
```
!python Lab2_Q2.py "/content/gdrive/My Drive/mr/empc.csv"

No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/Lab2_Q2.root.20240816.134851.532300
Running step 1 of 1...
job output is in /tmp/Lab2_Q2.root.20240816.134851.532300/output
Streaming final output from /tmp/Lab2_Q2.root.20240816.134851.532300/output...
null    "105,James,2027-11-10,M,55000,,1"
null    "106,Jennifer,1931-06-20,F,43000,105,4"
null    "107,Ahmad,1959-03-29,M,25000,106,4"
null    "103,Joyce,1962-07-31,F,25000,102,5"
null    "104,Ramesh,1952-09-15,M,38000,102,5"
null    "101,John,1955-01-09,M,30000,102,5"
null    "102,Franklin,1945-12-08,M,40000,105,5"
null    "108,Alicia,1958-07-19,F,25000,106,4"
Removing temp directory /tmp/Lab2_Q2.root.20240816.134851.532300...
```

Exercise #3: Let you yourself figure out a way a map-reduce based solution to compute moving average of time series data. There is book titled "Data Algorithms" [6].
A copy of the book is placed in shared dataset folder itself. Chapter 6 of this book discusses the computation of Moving average using map-reduce.
Refer related section for this purpose. Choose "Example 2: Time Series Data (URL Visits)" as data space

Let you use dataset https://www.kaggle.com/datasets/bobnau/daily-website-visitors and compute monthly moving average of website visitors (first time and repeat)

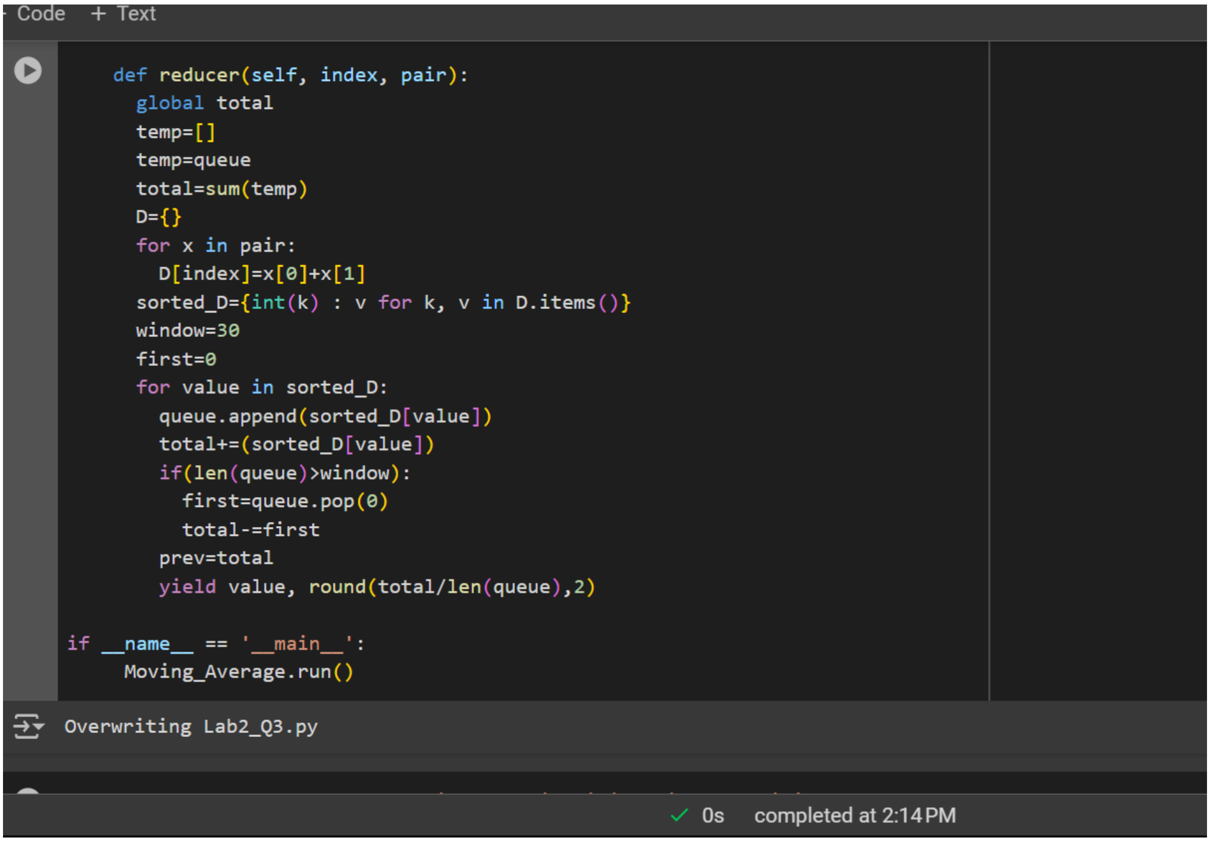
Code:

Mapper



```
import re
#import deque from collections
from mrjob.job import MRJob
from collections import OrderedDict
from queue import Queue
import re

total=0
global queue
queue=[]

class Moving_Average(MRJob):

    def mapper(self, _, line):
      record = re.split(',', line)
      l=len(record)

      # as the indices vary for each row from 9 to 12
      if(l==9):
        yield int(record[0]),(int(record[l - 2]) , int(record[l - 1]))
      elif(l==10):
        yield int(record[0]),(int(record[l - 2]) , int(record[l - 1]))
      elif(l==11):
        yield int(record[0]),(int(record[l - 3][1] + record[l - 2][:3]),int(record[l - 1]))
      elif(l==12):
        yield int(record[0]),(int(record[l - 4][1] + record[l - 3][:3]),int( record[l-2][1] + record[l - 1][:3]))
```

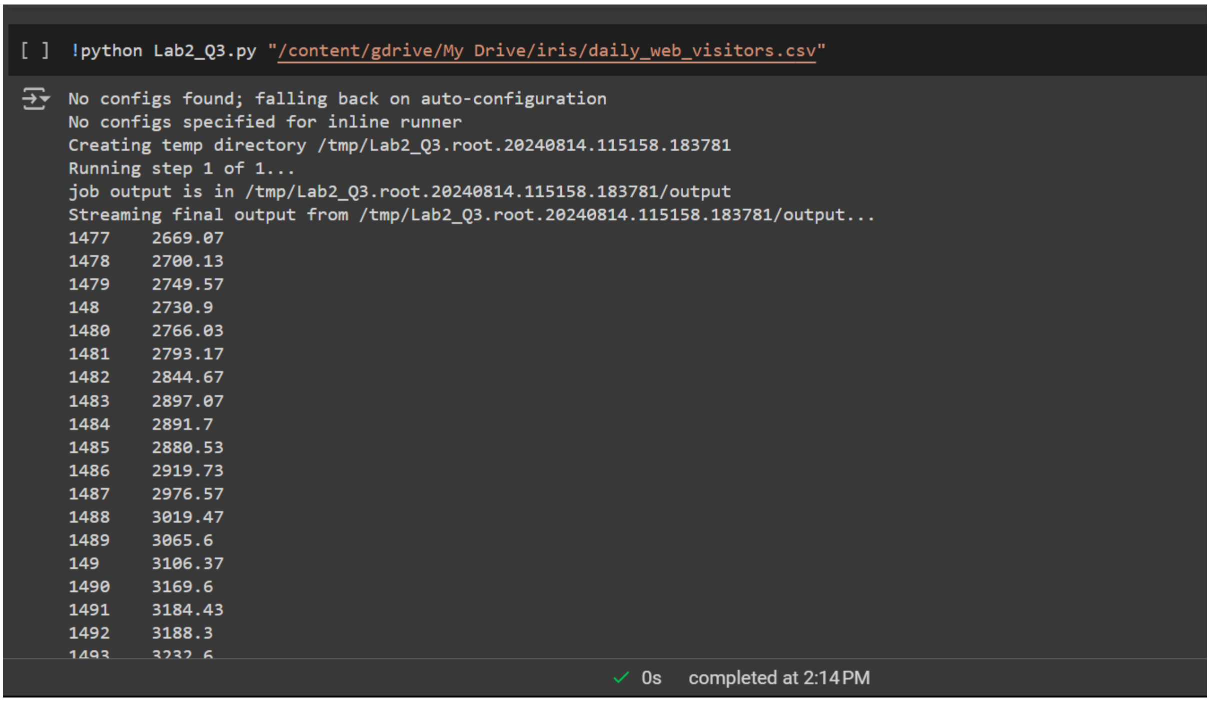
Reducer



```
    def reducer(self, index, pair):
      global total
      temp=[]
      temp=queue
      total=sum(temp)
      D={}
      for x in pair:
        D[index]=x[0]+x[1]
      sorted_D={int(k) : v for k, v in D.items()}
      window=30
      first=0
      for value in sorted_D:
        queue.append(sorted_D[value])
        total+=(sorted_D[value])
        if(len(queue)>window):
          first=queue.pop(0)
          total-=first
        prev=total
        yield value, round(total/len(queue),2)

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

```
Overwriting Lab2_Q3.py
```

Output:



```
!python Lab2_Q3.py "/content/gdrive/My Drive/iris/daily_web_visitors.csv"
```

```
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/Lab2_Q3.root.20240814.115158.183781
Running step 1 of 1...
job output is in /tmp/Lab2_Q3.root.20240814.115158.183781/output
Streaming final output from /tmp/Lab2_Q3.root.20240814.115158.183781/output...
1477    2669.07
1478    2700.13
1479    2749.57
148     2730.9
1480    2766.03
1481    2793.17
1482    2844.67
1483    2897.07
1484    2891.7
1485    2880.53
1486    2919.73
1487    2976.57
1488    3019.47
1489    3065.6
149     3106.37
1490    3169.6
1491    3184.43
1492    3188.3
1493    3232.6
```

*Note: In spite of my several attempts I was unable to sort the values on the basis of integer values. It was sorted lexicographically.