

5)WordCount.py : OUTPUT

```
"hadoop"      1
"how"         2
"individual"   1
"mrjob"       1
"oriented"    1
"python"      1
"reduce"      1
"when"        1
"available"   1
"combine"     1
"following"   1
"in"          1
"is"          2
"job"         4
"more"        2
"or"          2
"reference"    1
"submitted"   1
"task"        2
"uploaded"    1
"will"        1
"within"      1
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20231002.031149.100199...
Removing temp directory /tmp/WordCount.hadoop.20231002.031149.100199...
[hadoop@ip-172-31-27-113 ~]$
```

7)WordCount2.py :

PROGRAM :

```
from mrjob.job import MRJob
import re
```

```
PATTERN = re.compile(r"[\w']+")
```

```
class WordGroupCounter(MRJob):
```

```
    def mapper(self, _, record):
        for term in PATTERN.findall(record):
            group_key = "a_to_n" if 'a' <= term[0].lower() <= 'n' else "other"
            yield group_key, 1
```

```
    def combiner(self, group_key, totals):
        yield group_key, sum(totals)
```

```
    def reducer(self, group_key, totals):
        yield group_key, sum(totals)
```

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

OUTPUT

```
hadoop@ip-172-31-17-14:~$
Total time spent by all maps in occupied slots (ms)=10574016
Total time spent by all reduce tasks (ms)=71905
Total time spent by all reduces in occupied slots (ms)=6902880
Total vcore-milliseconds taken by all map tasks=220292
Total vcore-milliseconds taken by all reduce tasks=71905
Map-Reduce Framework
  CPU time spent (ms)=24740
  Combine input records=95
  Combine output records=6
  Failed Shuffles=0
  GC time elapsed (ms)=4616
  Input split bytes=1500
  Map input records=6
  Map output bytes=999
  Map output materialized bytes=1040
  Map output records=95
  Merged Map outputs=60
  Peak Map Physical memory (bytes)=552443904
  Peak Map Virtual memory (bytes)=3112923136
  Peak Reduce Physical memory (bytes)=301789184
  Peak Reduce Virtual memory (bytes)=4445618176
  Physical memory (bytes) snapshot=7296831488
  Reduce input groups=2
  Reduce input records=6
  Reduce output records=2
  Reduce shuffle bytes=1040
  Shuffled Maps =60
  Spilled Records=12
  Total committed heap usage (bytes)=6362759168
  Virtual memory (bytes) snapshot=59149344768
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
Job output is in hdfs:///user/hadoop/tmp/mrjob/wordCount2.hadoop.20231001.032914.735901/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/wordCount2.hadoop.20231001.032914.735901/output...
"cat" 49
"other" 46
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/wordCount2.hadoop.20231001.032914.735901...
Removing temp directory /tmp/wordCount2.hadoop.20231001.032914.735901...
[hadoop@ip-172-31-17-14 ~]$
```

9)Salaries.py : OUTPUT

```
"TRAFFIC INVESTIGATOR II" 7
"TRANSPORTATION ASSOC II" 9
"TRANSPORTATION SAFETY SUPERVIS" 1
"TREE SERVICE SUPV I" 2
"Transit Services Administrator" 1
"Transportation Enforcemt Sup II" 3
"UTILITIES INSTALLER REPAIR III" 47
"UTILITIES INSTALLER REPAIR SII" 15
"WATER SERVICE INSPECTOR" 4
"WATER SERVICE REPRESENTATIVE" 12
"WEB DEVELOPER" 1
"Web Chief of Engineering" 1
"Waste Water Maint Mgr Instrum" 1
"Waste Water Opns Tech II Pump" 10
"Waste Water Tech Supv I Pump" 6
"Water Systems Treatment Manage" 1
"Water Systems Treatment Supv" 2
"ZONING APPEALS ADVISOR BMZA" 1
"ZONING EXAMINER I" 2
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/salaries.hadoop.20231002.032042.219426...
Removing temp directory /tmp/salaries.hadoop.20231002.032042.219426...
[hadoop@ip-172-31-27-113 ~]$
```

11)Salaries2.py :

PROGRAM :

```
from mrjob.job import MRJob
```

```
class MRSalaryClassification(MRJob):
```

```
    def mapper(self, _, record_line):
```

```

try:
    (employee_name, position_title, org_id, organization, join_date, yearly_pay, total_pay) =
record_line.split('\t')

    processed_salary = float(yearly_pay.replace(",","").replace("$",""))

    if processed_salary >= 100000:
        yield "High", 1
    elif 50000 <= processed_salary < 100000:
        yield "Medium", 1
    else:
        yield "Low", 1
except Exception as processing_error:
    print(f"Error processing record: {record_line}. Error: {processing_error}")

def combiner(self, salary_category, category_counts):
    yield salary_category, sum(category_counts)

def reducer(self, salary_category, category_counts):
    yield salary_category, sum(category_counts)

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

```

OUTPUT :

```

Total time spent by all reduce tasks (ms)=54357
Total time spent by all reduces in occupied slots (ms)=5218272
Total vcore-milliseconds taken by all map tasks=223330
Total vcore-milliseconds taken by all reduce tasks=54357

Map-Reduce Framework
CPU time spent (ms)=27200
Combine input records=13818
Combine output records=36
Failed Shuffles=0
GC time elapsed (ms)=4519
Input split bytes=1584
Map input records=13818
Map output bytes=129922
Map output materialized bytes=1428
Map output records=13818
Merged Map outputs=60
Peak Map Physical memory (bytes)=546914304
Peak Map Virtual memory (bytes)=3146883072
Peak Reduce Physical memory (bytes)=332603392
Peak Reduce Virtual memory (bytes)=4438765568
Physical memory (bytes) snapshot=7495278592
Reduce input groups=3
Reduce input records=36
Reduce output records=3
Reduce shuffle bytes=1428
Shuffled Maps =60
Spilled Records=72
Total committed heap usage (bytes)=6601310208
Virtual memory (bytes) snapshot=59156664320

Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0

job output is in hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20231001.045636.525731/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20231001.045636.525731/output...
"Low"      7064
"High"     442
"Medium"   6312
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20231001.045636.525731...
Removing temp directory /tmp/Salaries.hadoop.20231001.045636.525731...
[hadoop@tp-172-31-22-212 ~]$

```

13) program to perform the task of outputting a count of the number of movies each user (identified via their user id) reviewed.

PROGRAM :

```
from mrjob.job import MRJob
from mrjob.step import MRStep

class MovieReview(MRJob):

    def steps(self):
        return [MRStep(mapper=self.distribute_data, reducer=self.collate_data)]

    def distribute_data(self, _, entry):
        user_id, film_id, _, _ = entry.split(',')
        yield user_id, 1

    def collate_data(self, user_id, occurrences):
        yield user_id, sum(occurrences)

if __name__ == '__main__':
    MovieReview(args=[data_source]).run()
```

OUTPUT :

```
"614" 99
"619" 43
"623" 103
"628" 87
"63" 97
"632" 39
"637" 25
"641" 140
"646" 169
"650" 29
"655" 105
"664" 519
"669" 37
"68" 123
"7" 88
"72" 191
"77" 315
"81" 160
"86" 190
"90" 50
"95" 299
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/user.hadoop.20231002.030434.866382...
Removing temp directory /tmp/user.hadoop.20231002.030434.866382...
[hadoop@ip-172-31-27-113 ~]$ |
```