ASSIGNMENT-03

1)We had modified the WordCount.py as WordCount2.py:

Program:

```
from mrjob.job import MRJob
import re

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

class wordcountmodified(MRJob):

    def mapper(self, _, line):
        for w in WORD_RE.findall(line):
            if w[0].lower()>='a' and w[0].lower()<='n':
                yield "a_to_n", 1
            else:
                yield "Other", 1

    def combiner(self, w, counts):
        yield w, sum(counts)

    def reducer(self, w, counts):
        yield w, sum(counts)


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

Executing Command:

```
[hadoop@ip-172-31-77-91 ~]$ python WordCount2.py -r hadoop hdfs:///user/hadoop/w.data
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.10.1
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
Creating temp directory /tmp/WordCount2.hadoop.20220129.194634.426609
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20220129.194634.426609/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20220129.194634.426609/files/
Running step 1 of 1...
  SLF4J: Class path contains multiple SLF4J bindings.
  SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
  SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
  SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
  SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-2.jar] /tmp/streamjob3317383566747843679.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-77-91.ec2.internal/172.31.77.91:8032
  Connecting to Application History server at ip-172-31-77-91.ec2.internal/172.31.77.91:10200
  Connecting to ResourceManager at ip-172-31-77-91.ec2.internal/172.31.77.91:8032
  Connecting to Application History server at ip-172-31-77-91.ec2.internal/172.31.77.91:10200
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 1
  number of splits:4
  Submitting tokens for job: job_1643481895791_0003
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
  Adding resource type - name = vcores, units = , type = COUNTABLE
  Submitted application application_1643481895791_0003
  The url to track the job: http://ip-172-31-77-91.ec2.internal:20888/proxy/application_1643481895791_0003/
  Running job: job_1643481895791_0003
  Job job_1643481895791_0003 running in uber mode : false
   map 0% reduce 0%
   map 50% reduce 0%
   map 75% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1643481895791_0003 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20220129.194634.426609/output
Counters: 50
        File Input Format Counters
                Bytes Read=1320
        File Output Format Counters
                Bytes Written=23
        File System Counters
                FILE: Number of bytes read=76
                FILE: Number of bytes written=1126650
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1764
                HDFS: Number of bytes written=23
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=15
                HDFS: Number of write operations=2
        Job Counters
                Data-local map tasks=4
                Killed map tasks=1
                Launched map tasks=4
```

Result:

```
STDERR: SLF4J: Class path contains multiple SLF4J bindings.
STDERR: SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
STDERR: SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
STDERR: SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
STDERR: SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
"Other" 46
"a_to_n"        49
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20220129.194634.426609...
Removing temp directory /tmp/WordCount2.hadoop.20220129.194634.426609...
[hadoop@ip-172-31-77-91 ~]$
```

2)

We had modified the Salaries.py to Salaries2.py:

Program:

```
from mrjob.job import MRJob

class salariesmodified(MRJob):

    def mapper(self, _, line):
        (name,jobTitle,agencyID,agency,hireDate,annualSalary,grossPay) = line.split('\t')
        if float(annualSalary) >= float(100000.00):
                yield "High", 1
        elif float(annualSalary) >= float(50000.00) and float(annualSalary) < float(100000.00):
                yield "Medium", 1
        else:
                yield "Low", 1

    def combiner(self, jobTitle, counts):
        yield jobTitle, sum(counts)

    def reducer(self, jobTitle, counts):
        yield jobTitle, sum(counts)


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

Executing Commands:

```
[hadoop@ip-172-31-77-91 ~]$ python Salaries2.py -r hadoop hdfs:///user/hadoop/Salaries.tsv
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.10.1
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
Creating temp directory /tmp/Salaries2.hadoop.20220129.200033.048061
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20220129.200033.048061/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20220129.200033.048061/files/
Running step 1 of 1...
  SLF4J: Class path contains multiple SLF4J bindings.
  SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
  SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
  SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
  SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-2.jar] /tmp/streamjob7880116301549984832.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-77-91.ec2.internal/172.31.77.91:8032
  Connecting to Application History server at ip-172-31-77-91.ec2.internal/172.31.77.91:10200
  Connecting to ResourceManager at ip-172-31-77-91.ec2.internal/172.31.77.91:8032
  Connecting to Application History server at ip-172-31-77-91.ec2.internal/172.31.77.91:10200
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 1
  number of splits:4
  Submitting tokens for job: job_1643481895791_0004
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
  Adding resource type - name = vcores, units = , type = COUNTABLE
  Submitted application application_1643481895791_0004
  The url to track the job: http://ip-172-31-77-91.ec2.internal:20888/proxy/application_1643481895791_0004/
  Running job: job_1643481895791_0004
  Job job_1643481895791_0004 running in uber mode : false
   map 0% reduce 0%
   map 50% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1643481895791_0004 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20220129.200033.048061/output
Counters: 50
        File Input Format Counters
                Bytes Read=1564110
        File Output Format Counters
                Bytes Written=36
        File System Counters
                FILE: Number of bytes read=116
                FILE: Number of bytes written=1126732
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1564578
                HDFS: Number of bytes written=36
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=15
                HDFS: Number of write operations=2
        Job Counters
                Data-local map tasks=4
                Killed map tasks=1
                Launched map tasks=4
                Launched reduce tasks=1
```

Result:

```
STDERR: SLF4J: Class path contains multiple SLF4J bindings.
STDERR: SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
STDERR: SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
STDERR: SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
STDERR: SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
"High"  442
"Low"   7064
"Medium"        6312
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20220129.200033.048061...
Removing temp directory /tmp/Salaries2.hadoop.20220129.200033.048061...
[hadoop@ip-172-31-77-91 ~]$
```

3) We had created a python file named as Movies.py:

Program:

```
from mrjob.job import MRJob

class moviesprogram(MRJob):

    def mapper(self, _, line):
        (userID,movieID,rating,timestamp) = line.split(',')
        yield userID, movieID

    def combiner(self, userID, movies):
        yield userID, len(list(movies))

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


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

Executing Commands:

```
[hadoop@ip-172-31-77-91 ~]$ python Movies.py -r hadoop hdfs:///user/hadoop/u.data
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.10.1
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
Creating temp directory /tmp/Movies.hadoop.20220129.201943.147684
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/Movies.hadoop.20220129.201943.147684/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/Movies.hadoop.20220129.201943.147684/files/
Running step 1 of 1...
  SLF4J: Class path contains multiple SLF4J bindings.
  SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
  SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
  SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
  SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-2.jar] /tmp/streamjob5996417966483771297.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-77-91.ec2.internal/172.31.77.91:8032
  Connecting to Application History server at ip-172-31-77-91.ec2.internal/172.31.77.91:10200
  Connecting to ResourceManager at ip-172-31-77-91.ec2.internal/172.31.77.91:8032
  Connecting to Application History server at ip-172-31-77-91.ec2.internal/172.31.77.91:10200
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 1
  number of splits:4
  Submitting tokens for job: job_1643481895791_0005
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
  Adding resource type - name = vcores, units = , type = COUNTABLE
  Submitted application application_1643481895791_0005
  The url to track the job: http://ip-172-31-77-91.ec2.internal:20888/proxy/application_1643481895791_0005/
  Running job: job_1643481895791_0005
  Job job_1643481895791_0005 running in uber mode : false
   map 0% reduce 0%
   map 50% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1643481895791_0005 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/Movies.hadoop.20220129.201943.147684/output
Counters: 50
        File Input Format Counters
                Bytes Read=2575317
        File Output Format Counters
                Bytes Written=6204
        File System Counters
                FILE: Number of bytes read=4636
                FILE: Number of bytes written=1135722
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2575761
                HDFS: Number of bytes written=6204
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=15
                HDFS: Number of write operations=2
        Job Counters
                Data-local map tasks=4
                Killed map tasks=1
                Launched map tasks=4
                Launched reduce tasks=1
```

Results:

```
job output is in hdfs:///user/hadoop/tmp/mrjob/Movies.hadoop.20220129.201943.147684/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Movies.hadoop.20220129.201943.147684/output...
"1"     20
"10"    46
"100"   25
"101"   55
"102"   678
"103"   94
"104"   76
"105"   525
"106"   45
"107"   32
"108"   31
"109"   23
"11"    38
"110"   120
"111"   341
"112"   21
"113"   27
"114"   25
"115"   41
"116"   25
"117"   55
"118"   189
"119"   641
"12"    61
"120"   138
"121"   80
"122"   40
"123"   33
"124"   85
"125"   210
"126"   64
"127"   21
"128"   323
"129"   26
"13"    53
"130"   375
"131"   44
"132"   94
"133"   178
"134"   311
"135"   22
"136"   50
"137"   80
"138"   81
"139"   68
"14"    20
"140"   46
"141"   31
"142"   61
"143"   77
"144"   41
"145"   38
"146"   73
"147"   38
"148"   132
"149"   231
"15"    1700
"150"   413
"151"   64
"152"   218
"153"   51
```