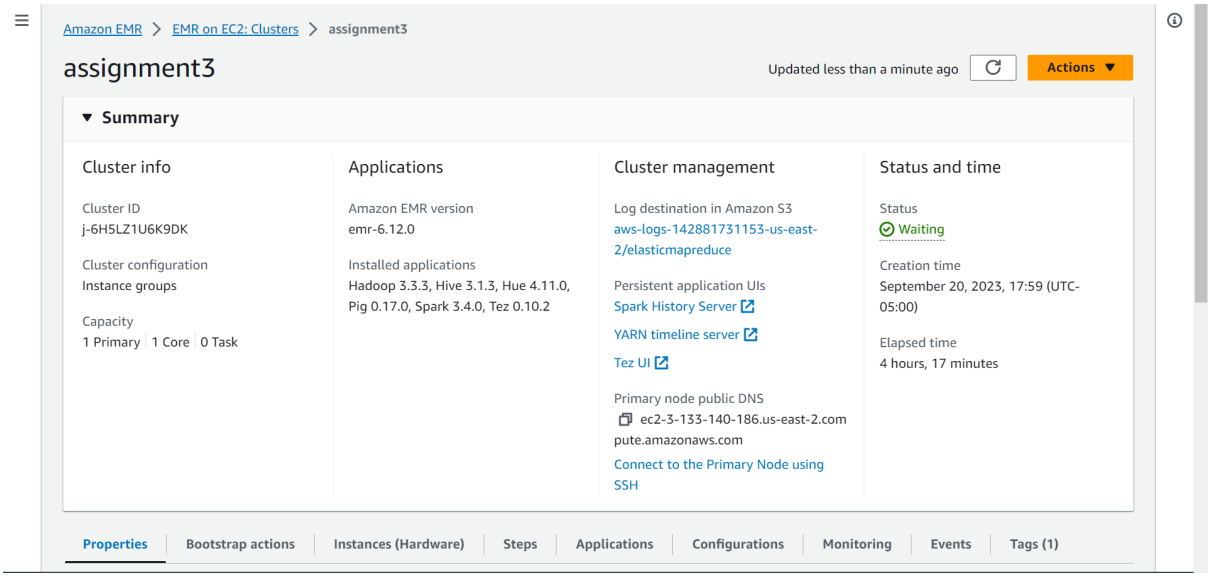
BigData
Assignment-3

4.Creating the cluster.



EMR connection

```
suman@Sumanth MINGW64 ~
$ cd downloads

suman@Sumanth MINGW64 ~/downloads
$ cd assign3-sai

suman@Sumanth MINGW64 ~/downloads/assign3-sai
$ chmod 400 emrkey.pem

suman@Sumanth MINGW64 ~/downloads/assign3-sai
$ ssh -i emrkey.pem hadoop@ec2-3-133-140-186.us-east-2.compute.amazonaws.com
The authenticity of host 'ec2-3-133-140-186.us-east-2.compute.amazonaws.com (3.133.140.186)' can't be established.
ED25519 key fingerprint is SHA256:INpytm+Poao7T30OcVXEdNDqMbt24u9QQo8npj5cqh8.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-3-133-140-186.us-east-2.compute.amazonaws.com' (ED25519) to the list of known hosts.

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/

EEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R::::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE:::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR
```

5.The mrjob library has been set up on the EMR primary node.

```
[hadoop@ip-172-31-3-230 ~]$ sudo /usr/bin/pip3.7 install mrjob[aws]
WARNING: Running pip install with root privileges is generally not a good idea. Try `pip3.7 install --user` instead.
Collecting mrjob[aws]
  Downloading mrjob-0.7.4-py2.py3-none-any.whl (439 kB)
     |████████████████████████████████| 439 kB 7.0 MB/s
Requirement already satisfied: PyYAML>=3.10 in /usr/local/lib64/python3.7/site-packages (from mrjob[aws]) (5.4.1)
Collecting botocore>=1.13.26; extra == "aws"
  Downloading botocore-1.31.52-py3-none-any.whl (11.2 MB)
     |████████████████████████████████| 11.2 MB 35.9 MB/s
Collecting boto3>=1.10.0; extra == "aws"
  Downloading boto3-1.28.52-py3-none-any.whl (135 kB)
     |████████████████████████████████| 135 kB 51.8 MB/s
Requirement already satisfied: jmespath<2.0.0,>=0.7.1 in /usr/local/lib/python3.7/site-packages (from botocore>=1.13.26; extra == "aws"->mrjob[aws]) (1.0.1)
Collecting urllib3<1.27,>=1.25.4
  Downloading urllib3-1.26.16-py2.py3-none-any.whl (143 kB)
     |████████████████████████████████| 143 kB 52.3 MB/s
Collecting python-dateutil<3.0.0,>=2.1
  Downloading python_dateutil-2.8.2-py2.py3-none-any.whl (247 kB)
     |████████████████████████████████| 247 kB 46.4 MB/s
Collecting s3transfer<0.7.0,>=0.6.0
  Downloading s3transfer-0.6.2-py3-none-any.whl (79 kB)
     |████████████████████████████████| 79 kB 11.9 MB/s
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/site-packages (from python-dateutil<3.0.0,>=2.1->botocore>=1.13.26; extra == "aws"->mrjob[aws]) (1.13.0)
Installing collected packages: urllib3, python-dateutil, botocore, s3transfer, boto3, mrjob
  WARNING: The scripts mrjob, mrjob-3 and mrjob-3.7 are installed in '/usr/local/bin' which is not on PATH.
  Consider adding this directory to PATH or, if you prefer to suppress this warning, use --no-warn-script-location.
Successfully installed boto3-1.28.52 botocore-1.31.52 mrjob-0.7.4 python-dateutil-2.8.2 s3transfer-0.6.2 urllib3-1.26.16
[hadoop@ip-172-31-3-230 ~]$
```

6.

Step 1: Firstly Move WordCount.py to home/Hadoop.

```
suman@Sumanth MINGW64 ~/downloads/assign3-sai
$ scp -i emrkey.pem WordCount2.py hadoop@ec2-3-133-140-186.us-east-2.compute.ama
zonaws.com:/home/hadoop
WordCount2.py                                 100%  499     7.9KB/s   00:00
```

Next,Move w.data to home/Hadoop

```
suman@Sumanth MINGW64 ~/downloads/assign3-sai
$ scp -i emrkey.pem w.data hadoop@ec2-3-133-140-186.us-east-2.compute.amazonaws.
com:/home/hadoop
w.data                                        100%  528    15.0KB/s   00:00
```

Step 2 & 3 : Now,Move w.data to user/hadoop

```
[hadoop@ip-172-31-3-230 ~]$ hadoop fs -copyFromLocal /home/hadoop/w.data /user/hadoop/w.data
[hadoop@ip-172-31-3-230 ~]$ hadoop fs -ls /user/hadoop/
Found 1 items
-rw-r--r--   1 hadoop hdfsadmingroup        528 2023-09-20 23:38 /user/hadoop/w.data
[hadoop@ip-172-31-3-230 ~]$
```

Step 4: Then Run the WordCount.py

```
[hadoop@ip-172-31-3-230 ~]$ python WordCount.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 3.3.3
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/WordCount.hadoop.20230920.235515.110716
Uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20230920.235515.110716/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20230920.235515.110716/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-4.jar] /tmp/streamjob8436192907519237160.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-3-230.us-east-2.compute.internal/172.31.3.230:8032
  Connecting to Application History server at ip-172-31-3-230.us-east-2.compute.internal/172.31.3.230:10200
  Connecting to ResourceManager at ip-172-31-3-230.us-east-2.compute.internal/172.31.3.230:8032
  Connecting to Application History server at ip-172-31-3-230.us-east-2.compute.internal/172.31.3.230:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1695251332922_0002
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 1
  number of splits:8
  Submitting tokens for job: job_1695251332922_0002
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1695251332922_0002
  The url to track the job: http://ip-172-31-3-230.us-east-2.compute.internal:20888/proxy/application_1695251332922_0002/
  Running job: job_1695251332922_0002
  Job job_1695251332922_0002 running in uber mode : false
   map 0% reduce 0%
   map 13% reduce 0%
   map 25% reduce 0%
   map 50% reduce 0%
   map 63% reduce 0%
   map 75% reduce 0%
   map 88% reduce 0%
   map 100% reduce 0%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1695251332922_0002 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20230920.235515.110716/output
Counters: 55
        File Input Format Counters
                Bytes Read=2376
        File Output Format Counters
                Bytes Written=652
        File System Counters
                FILE: Number of bytes read=751
                FILE: Number of bytes written=3260337
                FILE: Number of large read operations=0
```

This is the Output for WordCount.py

```
        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/WordCount.hadoop.20230920.235515.110716/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20230920.235515.110716/output...
"an"    1
"are"   1
"available"     1
"by"    1
"combine"       1
"defined"       1
"dependencies"  1
"for"   1
"hadoop"        1
"job"   4
"machine"       1
"map"   1
"more"  2
"of"    1
"or"    2
"our"   1
"python"        1
"script"        1
"task"  2
"the"   4
"within"        1
"a"     3
"all"   1
"and"   1
"be"    3
"do"    1
"either"        1
"first" 1
"following"     1
"how"   2
"is"    2
"must"  1
"nodes" 1
"oriented"      1
"reduce"        1
"reference"     1
"sections"      1
"that"  1
"two"   1
"versions"      1
"well"  1
"your"  5
"as"    4
"cluster"       2
"contained"     1
```

```
"cluster"       2
"contained"     1
"executed"      1
"explains"      1
"file"  2
"in"    1
"individual"    1
"mrjob" 1
"on"    4
"program"       1
"run"   1
"runners"       1
"second"        1
"see"   1
"submitted"     1
"things"        1
"those" 1
"to"    3
"uploaded"      1
"when"  1
"will"  1
"writing"       2
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/WordCount.hadoop.20230920.235515.110716...
Removing temp directory /tmp/WordCount.hadoop.20230920.235515.110716...
[hadoop@ip-172-31-3-230 ~]$
```

Step 5: Now Export the WordCount2.py to home/hadoop

```
suman@Sumanth MINGW64 ~/downloads/assign3-sai
$ scp -i emrkey.pem WordCount2.py hadoop@ec2-3-133-140-186.us-east-2.compute.ama
zonaws.com:/home/hadoop
WordCount2.py                                 100%  499     7.9KB/s   00:00
```

6) The updated Program of WordCount2.py

C: > Users > suman > Downloads > assign3-sai > WordCount2.py > ...

```python
from multiprocessing import Pool
import re
from collections import defaultdict

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

def count_words_in_text(text):
    word_counts = defaultdict(int)
    words = WORD_RE.findall(text)

    for word in words:
        if re.match(r'[a-n]', word[0]):
            word_counts['a_to_n'] += 1
        else:
            word_counts['other'] += 1

    return word_counts

def merge_counts(counts_list):
    total_counts = defaultdict(int)

    for counts in counts_list:
        for word, count in counts.items():
            total_counts[word] += count

    return total_counts

if __name__ == '__main__':
    with open('input.txt', 'r') as file:
        lines = file.readlines()

    with Pool() as pool:
        word_counts_list = pool.map(count_words_in_text, lines)
```

```python
    with Pool() as pool:
        word_counts_list = pool.map(count_words_in_text, lines)

    total_counts = merge_counts(word_counts_list)

    for word, count in total_counts.items():
        print(f'{word}: {count}')
```

Run WordCount2.py

```
[hadoop@ip-172-31-3-230 ~]$ 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 3.3.3
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.20230921.004550.299915
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20230921.004550.299915/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20230921.004550.299915/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-4.jar] /tmp/streamjob6597987648109307924.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-3-230.us-east-2.compute.internal/172.31.3.230:8032
  Connecting to Application History server at ip-172-31-3-230.us-east-2.compute.internal/172.31.3.230:10200
  Connecting to ResourceManager at ip-172-31-3-230.us-east-2.compute.internal/172.31.3.230:8032
  Connecting to Application History server at ip-172-31-3-230.us-east-2.compute.internal/172.31.3.230:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1695251332922_0006
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 1
  number of splits:8
  Submitting tokens for job: job_1695251332922_0006
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1695251332922_0006
  The url to track the job: http://ip-172-31-3-230.us-east-2.compute.internal:20888/proxy/application_1695251332922_0006/
  Running job: job_1695251332922_0006
  Job job_1695251332922_0006 running in uber mode : false
   map 0% reduce 0%
   map 13% reduce 0%
   map 63% reduce 0%
   map 75% reduce 0%
   map 100% reduce 0%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1695251332922_0006 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20230921.004550.299915/output
Counters: 55
        File Input Format Counters
                Bytes Read=2376
        File Output Format Counters
                Bytes Written=23
        File System Counters
                FILE: Number of bytes read=118
                FILE: Number of bytes written=3259104
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3376
```

The Output for WordCount2.py

```
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=23
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=39
                HDFS: Number of write operations=6
        Job Counters
                Data-local map tasks=8
                Killed map tasks=1
                Launched map tasks=8
                Launched reduce tasks=3
                Total megabyte-milliseconds taken by all map tasks=182733312
                Total megabyte-milliseconds taken by all reduce tasks=65369088
                Total time spent by all map tasks (ms)=118967
                Total time spent by all maps in occupied slots (ms)=5710416
                Total time spent by all reduce tasks (ms)=21279
                Total time spent by all reduces in occupied slots (ms)=2042784
                Total vcore-milliseconds taken by all map tasks=118967
                Total vcore-milliseconds taken by all reduce tasks=21279
        Map-Reduce Framework
                CPU time spent (ms)=19760
                Combine input records=95
                Combine output records=6
                Failed Shuffles=0
                GC time elapsed (ms)=2604
                Input split bytes=1000
                Map input records=6
                Map output bytes=996
                Map output materialized bytes=464
                Map output records=95
                Merged Map outputs=24
                Peak Map Physical memory (bytes)=551849984
                Peak Map Virtual memory (bytes)=3101724672
                Peak Reduce Physical memory (bytes)=329314304
                Peak Reduce Virtual memory (bytes)=4434558976
                Physical memory (bytes) snapshot=4961406976
                Reduce input groups=2
                Reduce input records=6
                Reduce output records=2
                Reduce shuffle bytes=464
                Shuffled Maps =24
                Spilled Records=12
                Total committed heap usage (bytes)=4599054336
                Virtual memory (bytes) snapshot=37924429824
        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.20230921.004550.299915/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/WordCount2.hadoop.20230921.004550.299915/output...
"a_to_n"        46
"other" 49
```

7.Now move the Salaries.py , Salaries.tsv , Salaries2.py to /home/hadoop

```
suman@Sumanth MINGW64 ~/downloads/assign3-sai
$ scp -i emrkey.pem Salaries.py hadoop@ec2-3-133-140-186.us-east-2.compute.amazonaws.com:/home/hadoop
Salaries.py                                                                   100%  411    12.8KB/s   00:00

suman@Sumanth MINGW64 ~/downloads/assign3-sai
$ scp -i emrkey.pem Salaries.tsv hadoop@ec2-3-133-140-186.us-east-2.compute.amazonaws.com:/home/hadoop
Salaries.tsv                                                                  100% 1502KB 337.4KB/s   00:04

suman@Sumanth MINGW64 ~/downloads/assign3-sai
$ scp -i emrkey.pem Salaries2.py hadoop@ec2-3-133-140-186.us-east-2.compute.amazonaws.com:/home/hadoop
Salaries2.py                                                                  100%  683    17.9KB/s   00:00
```

The files were obtained from Salaries.tsv.

```
[hadoop@ip-172-31-3-230 ~]$ hadoop fs -put /home/hadoop/Salaries.tsv hdfs:///user/Hadoop
[hadoop@ip-172-31-3-230 ~]$ hadoop fs -ls /user/hadoop
Found 5 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-09-21 00:40 /user/hadoop/WordCount2
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-09-20 23:49 /user/hadoop/dout
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-09-21 00:10 /user/hadoop/downloads
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-09-20 23:48 /user/hadoop/tmp
-rw-r--r--   1 hadoop hdfsadmingroup        528 2023-09-20 23:38 /user/hadoop/w.data
[hadoop@ip-172-31-3-230 ~]$
```

8) Executing the Salaries.py

```
6.
[hadoop@ip-172-31-3-230 ~]$ python Salaries.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 3.3.3
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/Salaries.hadoop.20230921.014728.924469
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20230921.014728.924469/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20230921.014728.924469/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-4.jar] /tmp/streamjob7528707264152856511.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-3-230.us-east-2.compute.internal/172.31.3.230:8032
  Connecting to Application History server at ip-172-31-3-230.us-east-2.compute.internal/172.31.3.230:10200
  Connecting to ResourceManager at ip-172-31-3-230.us-east-2.compute.internal/172.31.3.230:8032
  Connecting to Application History server at ip-172-31-3-230.us-east-2.compute.internal/172.31.3.230:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1695251332922_0008
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 1
  number of splits:8
  Submitting tokens for job: job_1695251332922_0008
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1695251332922_0008
  The url to track the job: http://ip-172-31-3-230.us-east-2.compute.internal:20888/proxy/application_1695251332922_0008/
  Running job: job_1695251332922_0008
  Job job_1695251332922_0008 running in uber mode : false
   map 0% reduce 0%
   map 75% reduce 0%
   map 100% reduce 0%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1695251332922_0008 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/Salaries.hadoop.20230921.014728.924469/output
Counters: 55
        File Input Format Counters
                Bytes Read=1567508
        File Output Format Counters
                Bytes Written=29260
        File System Counters
                FILE: Number of bytes read=27045
                FILE: Number of bytes written=3350196
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1568556
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=29260
                HDFS: Number of large read operations=0
```

This is the Output for Salaries.py

```
"RETIREMENT BENEFITS MANAGER"   1
"Recreation Leader II Elder Act"        1
"Research Analyst I"    1
"SAFETY ENFORCEMENT OFFICER I"  2
"SALES MANAGER" 3
"SCADA System Supervisor"       2
"SECRETARY III" 34
"SENIOR SOCIAL SERVICES COORDIN"        13
"SENIOR YOUTH DEVELOPMENT TECHN"        1
"SERV ASST LBRY"        50
"SERVICE AIDE II"       1
"SEWERLINE VIDEO INSPECTOR TECH"        4
"SHERIFF"       1
"SIGN FABRICATOR I"     2
"SIGN PAINTER II"       4
"SOCIAL PROG ADMINISTRATOR III" 1
"SOLID WASTE SUPERINTENDENT"    4
"SR COMPANION STIPEND HLTH"     143
"STATE LIBRARY RESOURCE CENTER" 3
"STATE'S ATTORNEY"      1
"STATISTICAL TRAFFIC ANALYST"   1
"STOREKEEPER I" 22
"STORES SUPERVISOR II"  2
"STREET MASON"  1
"SUPT CLEANING BOARDNG & GR MNT"        1
"SUPT COMMUNICATIONS/COMPUTER O"        1
"SUPT PLANS AND INSPECTIONS"    2
"SUPT TRAFFIC SIGNAL INSTALLATI"        1
"SUPV. OF BOARDING/GROUNDS MAIN"        1
"SURVEY COMPUTATION ANALYST"    1
"SURVEY TECHNICIAN II"  3
"SURVEY TECHNICIAN III" 1
"SWIMMING POOL ATTENDENT"       26
"SYSTEMS SUPERVISOR"    2
"Senior Fire Operations Aide"   2
"Solid Waste Asst Superintenden"        2
"Systems Analyst"       3
"TOWING LOT SUPERINTENDENT"     1
"TRACTOR TRAILER DRIVER"        5
"TRAFFIC INVESTIGATOR III"      2
"TREASURY ASSISTANT"    1
"TREASURY TECHNICIAN"   2
"Transportation Enforcemt Off I"        65
"Transportation Enforcmt Off II"        20
"Transportation Enforcmt Sup II"        3
"UTILITIES INSTALLER REPAIR III"        47
[hadoop@ip-172-31-3-230 ~]$
[hadoop@ip-172-31-3-230 ~]$
```

9) The updated code for Salaries2.py

C: > Users > suman > Downloads > assign3-sai > Salaries2.py > ...

```python
import pandas as pd

def classify_salary_range(annual_salary):
    annual_salary = float(annual_salary)
    if annual_salary >= 100000.0:
        return 'High'
    elif 50000.0 <= annual_salary <= 99999.99:
        return 'Medium'
    elif 0.0 <= annual_salary <= 49999.99:
        return 'Low'
    else:
        return 'Invalid'

if __name__ == '__main__':
    df = pd.read_csv('salaries.txt', sep='\t', header=None,
                     names=['name', 'jobTitle', 'agencyID', 'agency', 'hireDate', 'annualSalary', 'grossPay'])

    df['salary_range'] = df['annualSalary'].apply(classify_salary_range)
    salary_counts = df['salary_range'].value_counts().to_dict()

    for salary_range, count in salary_counts.items():
        print(f'{salary_range}: {count}')
```

10) Running the Salaries2.py

```
[hadoop@ip-172-31-3-230 ~]$ 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 3.3.3
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.20230921.021805.529450
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20230921.021805.529450/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20230921.021805.529450/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-4.jar] /tmp/streamjob7704840667057651238.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-3-230.us-east-2.compute.internal/172.31.3.230:8032
  Connecting to Application History server at ip-172-31-3-230.us-east-2.compute.internal/172.31.3.230:10200
  Connecting to ResourceManager at ip-172-31-3-230.us-east-2.compute.internal/172.31.3.230:8032
  Connecting to Application History server at ip-172-31-3-230.us-east-2.compute.internal/172.31.3.230:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1695251332922_0009
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 1
  number of splits:8
  Submitting tokens for job: job_1695251332922_0009
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1695251332922_0009
  The url to track the job: http://ip-172-31-3-230.us-east-2.compute.internal:20888/proxy/application_1695251332922_0009/
  Running job: job_1695251332922_0009
  Job job_1695251332922_0009 running in uber mode : false
   map 0% reduce 0%
   map 13% reduce 0%
   map 75% reduce 0%
   map 100% reduce 0%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1695251332922_0009 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20230921.021805.529450/output
Counters: 55
        File Input Format Counters
                Bytes Read=1567508
        File Output Format Counters
                Bytes Written=36
        File System Counters
                FILE: Number of bytes read=210
                FILE: Number of bytes written=3259343
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1568556
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=36
```

This is the Output for Salaries2.py

```
                HDFS: Number of read operations=39
                HDFS: Number of write operations=6
        Job Counters
                Data-local map tasks=8
                Killed map tasks=1
                Launched map tasks=8
                Launched reduce tasks=3
                Total megabyte-milliseconds taken by all map tasks=213828096
                Total megabyte-milliseconds taken by all reduce tasks=71344128
                Total time spent by all map tasks (ms)=139211
                Total time spent by all maps in occupied slots (ms)=6682128
                Total time spent by all reduce tasks (ms)=23224
                Total time spent by all reduces in occupied slots (ms)=2229504
                Total vcore-milliseconds taken by all map tasks=139211
                Total vcore-milliseconds taken by all reduce tasks=23224
        Map-Reduce Framework
                CPU time spent (ms)=25750
                Combine input records=13818
                Combine output records=24
                Failed Shuffles=0
                GC time elapsed (ms)=2701
                Input split bytes=1048
                Map input records=13818
                Map output bytes=129922
                Map output materialized bytes=696
                Map output records=13818
                Merged Map outputs=24
                Peak Map Physical memory (bytes)=548196352
                Peak Map Virtual memory (bytes)=3144253440
                Peak Reduce Physical memory (bytes)=276357120
                Peak Reduce Virtual memory (bytes)=4443521024
                Physical memory (bytes) snapshot=4951101440
                Reduce input groups=3
                Reduce input records=24
                Reduce output records=3
                Reduce shuffle bytes=696
                Shuffled Maps =24
                Spilled Records=48
                Total committed heap usage (bytes)=4406116352
                Virtual memory (bytes) snapshot=38002135040
        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/Salaries2.hadoop.20230921.021805.529450/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20230921.021805.529450/output...
"High"  442
"Low"   7064
"Medium"        6312
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/Salaries2.hadoop.20230921.021805.529450...
Removing temp directory /tmp/Salaries2.hadoop.20230921.021805.529450...
```

11) Moving u.data to /home/hadoop

```
suman@Sumanth MINGW64 ~/downloads/assign3-sai
$ scp -i emrkey.pem u.data hadoop@ec2-3-133-140-186.us-east-2.compute.amazonaws.com:/home/hadoop
u.data                                                                    100% 2381KB 416.7KB/s   00:05
```

```
[hadoop@ip-172-31-3-230 ~]$ hadoop fs -put /home/hadoop/u.data hdfs:///user/hadoop
[hadoop@ip-172-31-3-230 ~]$ hadoop fs -ls /user/hadoop
Found 8 items
-rw-r--r--   1 hadoop hdfsadmingroup        411 2023-09-21 01:27 /user/hadoop/Salaries.py
-rw-r--r--   1 hadoop hdfsadmingroup    1538148 2023-09-21 01:27 /user/hadoop/Salaries.tsv
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-09-21 00:40 /user/hadoop/WordCount2
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-09-20 23:49 /user/hadoop/dout
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-09-21 00:10 /user/hadoop/downloads
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-09-20 23:48 /user/hadoop/tmp
-rw-r--r--   1 hadoop hdfsadmingroup    2438233 2023-09-21 02:55 /user/hadoop/u.data
-rw-r--r--   1 hadoop hdfsadmingroup        528 2023-09-20 23:38 /user/hadoop/w.data
[hadoop@ip-172-31-3-230 ~]$
```

12) Move movies1.py to /home/hadoop

```
suman@Sumanth MINGW64 ~/downloads/assign3-sai
$ scp -i emrkey.pem Moviesrating.py hadoop@ec2-3-133-140-186.us-east-2.compute.amazonaws.com:/home/hadoop
Moviesrating.py                                   100%  409    10.8KB/s   00:00
```

Updated Code for movies1.py

C: > Users > suman > Downloads > assign3-sai > Moviesrating.py > ...

```python
from collections import defaultdict

if __name__ == '__main__':
    with open('movies.csv', 'r') as file:
        lines = file.readlines()

    user_counts = defaultdict(int)

    for line in lines:
        _, movie_id, _, _ = line.strip().split(',')
        user_counts[movie_id] += 1

    for movie_id, count in user_counts.items():
        print(f'Movie ID: {movie_id}, User Count: {count}')
```

Running the movies1.py

```
[hadoop@ip-172-31-3-230 ~]$ python Moviesrating.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 3.3.3
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/Moviesrating.hadoop.20230921.025708.455704
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/Moviesrating.hadoop.20230921.025708.455704/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/Moviesrating.hadoop.20230921.025708.455704/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-4.jar] /tmp/streamjob7309968150198101072.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-3-230.us-east-2.compute.internal/172.31.3.230:8032
  Connecting to Application History server at ip-172-31-3-230.us-east-2.compute.internal/172.31.3.230:10200
  Connecting to ResourceManager at ip-172-31-3-230.us-east-2.compute.internal/172.31.3.230:8032
  Connecting to Application History server at ip-172-31-3-230.us-east-2.compute.internal/172.31.3.230:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1695251332922_0010
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 1
  number of splits:8
  Submitting tokens for job: job_1695251332922_0010
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1695251332922_0010
  The url to track the job: http://ip-172-31-3-230.us-east-2.compute.internal:20888/proxy/application_1695251332922_0010/
  Running job: job_1695251332922_0010
  Job job_1695251332922_0010 running in uber mode : false
   map 0% reduce 0%
   map 13% reduce 0%
   map 75% reduce 0%
   map 88% reduce 0%
   map 100% reduce 0%
   map 100% reduce 33%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1695251332922_0010 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/Moviesrating.hadoop.20230921.025708.455704/output
Counters: 55
        File Input Format Counters
                Bytes Read=2597157
        File Output Format Counters
                Bytes Written=6204
        File System Counters
                FILE: Number of bytes read=5193
                FILE: Number of bytes written=3270493
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
```

This is the Output for movies1.py

```
"563"	158
"566"	22
"569"	85
"572"	106
"575"	547
"578"	34
"581"	49
"584"	193
"587"	504
"59"	78
"590"	89
"593"	70
"596"	487
"599"	192
"602"	129
"605"	437
"608"	296
"611"	35
"614"	99
"617"	75
"62"	53
"620"	172
"623"	103
"626"	150
"629"	34
"632"	39
"635"	22
"638"	20
"641"	140
"644"	39
"647"	150
"65"	27
"650"	29
"653"	51
"656"	128
"659"	142
"662"	58
"665"	434
"668"	20
"671"	115
"68"	123
"71"	23
"74"	49
"77"	315
"8"	116
"80"	37
"83"	161
"86"	190
"89"	66
"92"	123
"95"	299
"98"	71
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/Moviesrating.hadoop.20230921.025708.455704...
Removing temp directory /tmp/Moviesrating.hadoop.20230921.025708.455704...
```

Submitted by:
Sailavanya Narthu
A20516764
snarthu@hawk.iit.edu