Chaitanya Durgesh Nynavarapu
A20561894

ASSIGNMENT-3

BIG DATA (CSP 554)

1.

HaRD: A Heterogeneity-Aware Replica Deletion Algorithm for HDFS

Introduction

The Hadoop Distributed File System (HDFS) is widely used for storing large datasets reliably on clusters of commodity machines. Data replication in HDFS improves availability and performance but comes at the cost of increased storage usage. Recent studies have proposed dynamic replication management frameworks that adjust the replication factor based on data popularity. However, reducing the replication factor can lead to unbalanced data distribution, causing performance issues.

This paper identifies that existing replica deletion approaches, including Hadoop's default algorithm and the authors' previous WBRD (Workload-aware Balanced Replica Deletion) algorithm, perform sub-optimally on heterogeneous clusters. To address this limitation, the authors propose HaRD (Heterogeneity-aware Replica Deletion), a novel algorithm for deleting replicas in heterogeneous HDFS clusters.

Key Contributions

1. Extension of the formal definition of the replica deletion problem to heterogeneous clusters.
2. Proposal of HaRD, which considers nodes' processing capabilities when deleting replicas.
3. Implementation of HaRD on top of HDFS and extensive experiments on a 23-node heterogeneous cluster.

Methodology

HaRD aims to balance the ratio of block distribution to computing capabilities for each node. It determines a node's computing capability by calculating how many containers it can run simultaneously. This approach provides flexibility and minimal overhead.

The authors evaluated HaRD against Hadoop's default deletion algorithm and WBRD using various benchmarks:

- TestDFSIO
- Grep
- Terasort
- Concurrency test with TPC-H

Results

Key findings from the experiments include:

- HaRD reduced execution time by up to 60% compared to Hadoop and 17% compared to WBRD.
- HaRD achieved better data locality (85% vs 81% for WBRD and 73% for Hadoop).
- HaRD reduced network utilization by 6.9% compared to WBRD.
- Performance improvements were more significant under heavy loads with concurrent users.

Block Distribution Analysis

The authors analyzed the block distribution after reducing the replication factor from 10 to 3 using different deletion algorithms. They found that:

- Hadoop's deletion algorithm resulted in a skewed data distribution with high standard deviation.
- WBRD achieved an evenly balanced block distribution but did not consider node processing capabilities.
- HaRD stored more blocks on more powerful computers, creating three distinct groups in the block distribution corresponding to the three types of machines in the cluster

Performance Evaluation

The authors conducted experiments using TestDFSIO, Terasort, and Grep benchmarks with different replication factors:

- With RF=3, HaRD reduced average execution time by 7% for TestDFSIO, 6.1% for Terasort, and 9.4% for Grep compared to WBRD.
- With RF=1, HaRD's performance improvements were even more significant: 18.1% for TestDFSIO, 9.2% for Terasort, and 30.6% for Grep compared to WBRD

Concurrency Test

Using TPC-H Q6 with varying numbers of concurrent users (25 to 125), the authors found that:

- HaRD consistently outperformed both WBRD and Hadoop.
- Performance improvements became more significant as the number of concurrent users increased.
- With 125 concurrent users, HaRD reduced execution time by 17% compared to WBRD and 60% compared to Hadoop

Analysis

➢ For the 125 concurrent users test, the authors observed:
  - HaRD achieved 85% data locality, compared to 81% for WBRD and 73% for Hadoop.
  - HaRD reduced average network utilization by 6.9% compared to WBRD.
  - HaRD balanced network bandwidth usage across nodes, while Hadoop's usage was unbalanced due to "hot spots"

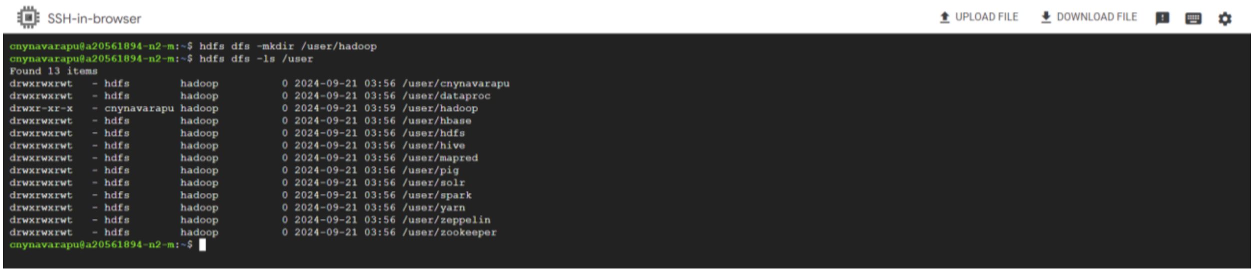
➢ The authors measured HaRD's implementation overhead:
  - HaRD introduced a 10.8 millisecond overhead for decreasing the replication factor from 10 to 3 for a 50 GB dataset.
  - The overhead scaled linearly with increasing data size and number of nodes, proving HaRD's scalability

Conclusion

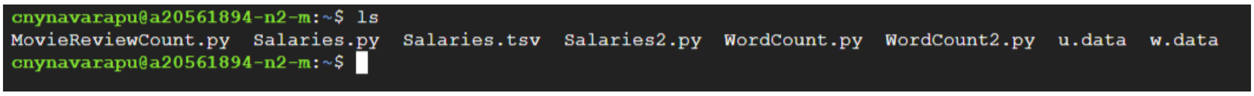
HaRD offers a cost-effective solution for replica deletion in heterogeneous Hadoop clusters, significantly improving performance over existing approaches. The authors suggest future work could involve developing an adaptive replication management framework using HaRD.

2.

- Created a "/user/hadoop"

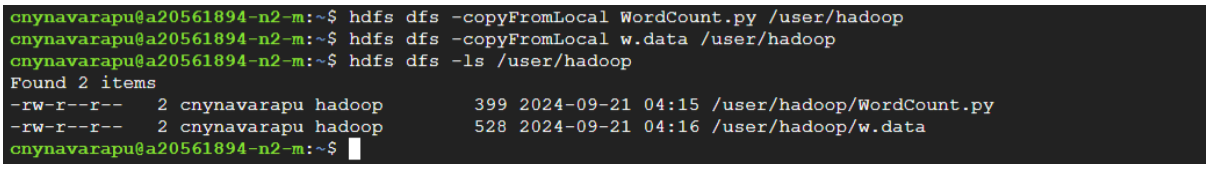


```
cnynavarapu@a20561894-n2-m:~$ hdfs dfs -mkdir /user/hadoop
cnynavarapu@a20561894-n2-m:~$ hdfs dfs -ls /user
Found 13 items
drwxrwxrwt   - hdfs        hadoop          0 2024-09-21 03:56 /user/cnynavarapu
drwxrwxrwt   - hdfs        hadoop          0 2024-09-21 03:56 /user/dataproc
drwxr-xr-x   - cnynavarapu hadoop          0 2024-09-21 03:59 /user/hadoop
drwxrwxrwt   - hdfs        hadoop          0 2024-09-21 03:56 /user/hbase
drwxrwxrwt   - hdfs        hadoop          0 2024-09-21 03:56 /user/hdfs
drwxrwxrwt   - hdfs        hadoop          0 2024-09-21 03:56 /user/hive
drwxrwxrwt   - hdfs        hadoop          0 2024-09-21 03:56 /user/mapred
drwxrwxrwt   - hdfs        hadoop          0 2024-09-21 03:56 /user/pig
drwxrwxrwt   - hdfs        hadoop          0 2024-09-21 03:56 /user/solr
drwxrwxrwt   - hdfs        hadoop          0 2024-09-21 03:56 /user/spark
drwxrwxrwt   - hdfs        hadoop          0 2024-09-21 03:56 /user/yarn
drwxrwxrwt   - hdfs        hadoop          0 2024-09-21 03:56 /user/zeppelin
drwxrwxrwt   - hdfs        hadoop          0 2024-09-21 03:56 /user/zookeeper
cnynavarapu@a20561894-n2-m:~$
```



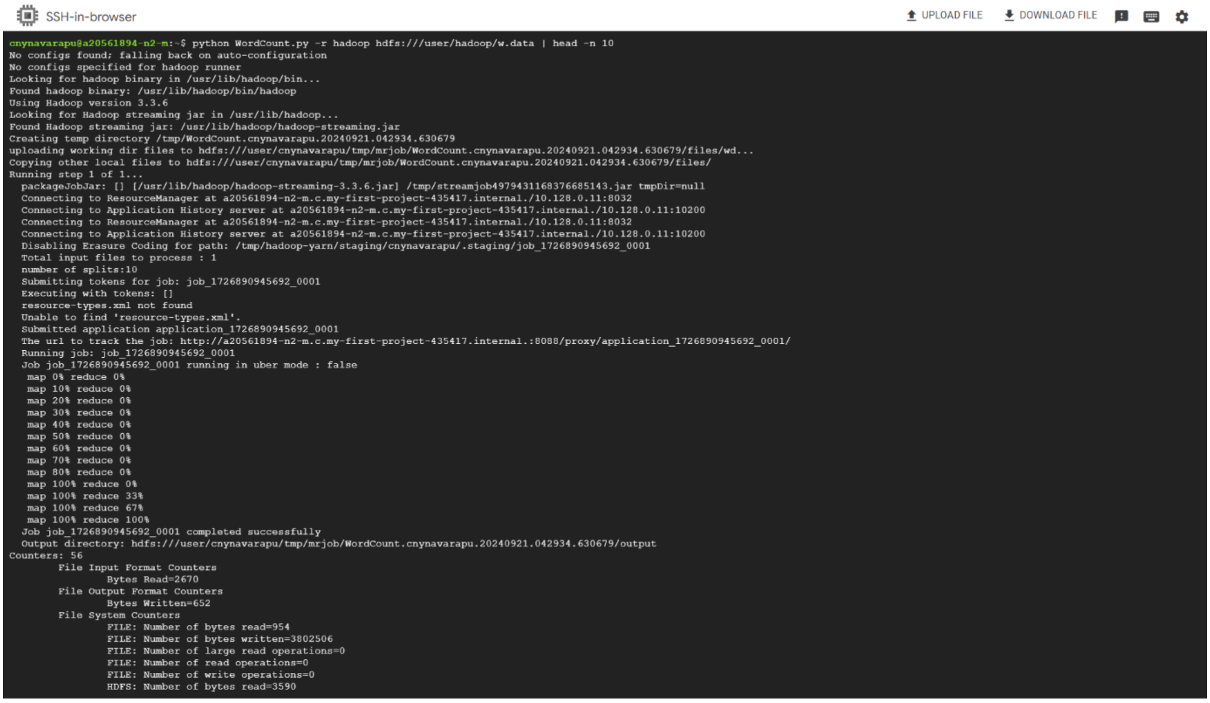
```
cnynavarapu@a20561894-n2-m:~$ ls
MovieReviewCount.py  Salaries.py  Salaries.tsv  Salaries2.py  WordCount.py  WordCount2.py  u.data  w.data
cnynavarapu@a20561894-n2-m:~$
```

- Loaded WordCount.py from local to “/user/hadoop” and have listed them

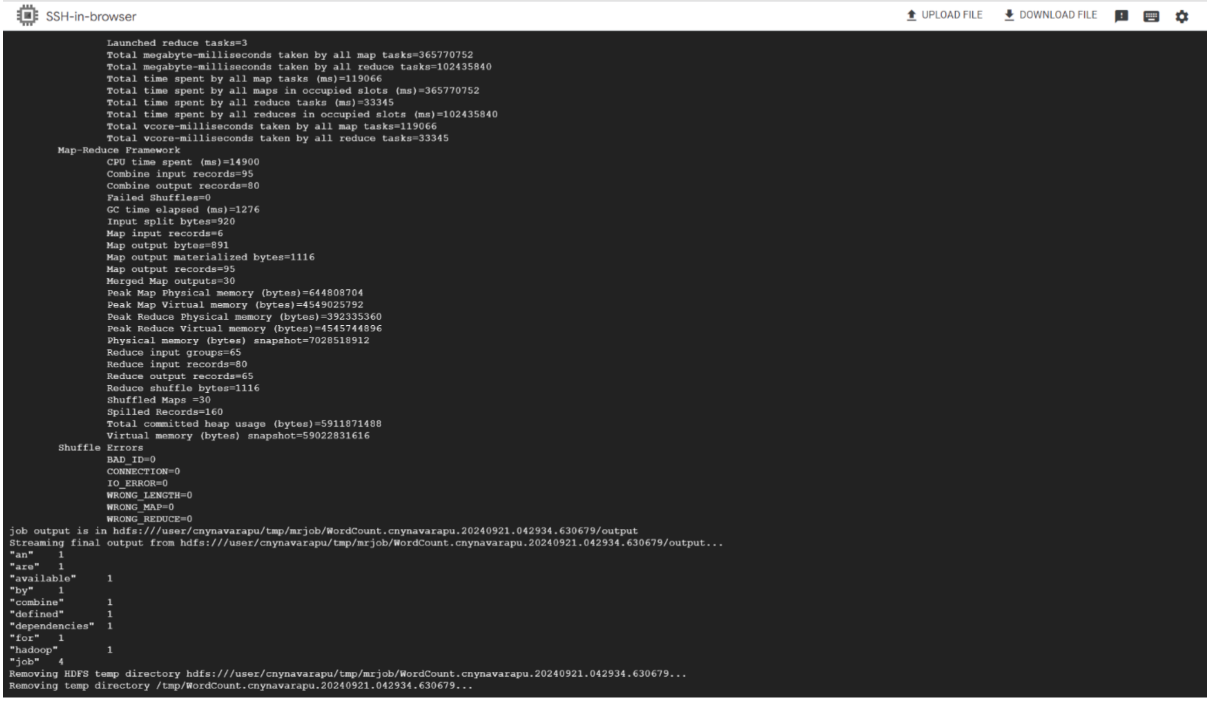


```
cnynavarapu@a20561894-n2-m:~$ hdfs dfs -copyFromLocal WordCount.py /user/hadoop
cnynavarapu@a20561894-n2-m:~$ hdfs dfs -copyFromLocal w.data /user/hadoop
cnynavarapu@a20561894-n2-m:~$ hdfs dfs -ls /user/hadoop
Found 2 items
-rw-r--r--   2 cnynavarapu hadoop        399 2024-09-21 04:15 /user/hadoop/WordCount.py
-rw-r--r--   2 cnynavarapu hadoop        528 2024-09-21 04:16 /user/hadoop/w.data
cnynavarapu@a20561894-n2-m:~$ 
```

- Performed given operation, and tested WordCount.py to print 10 head values

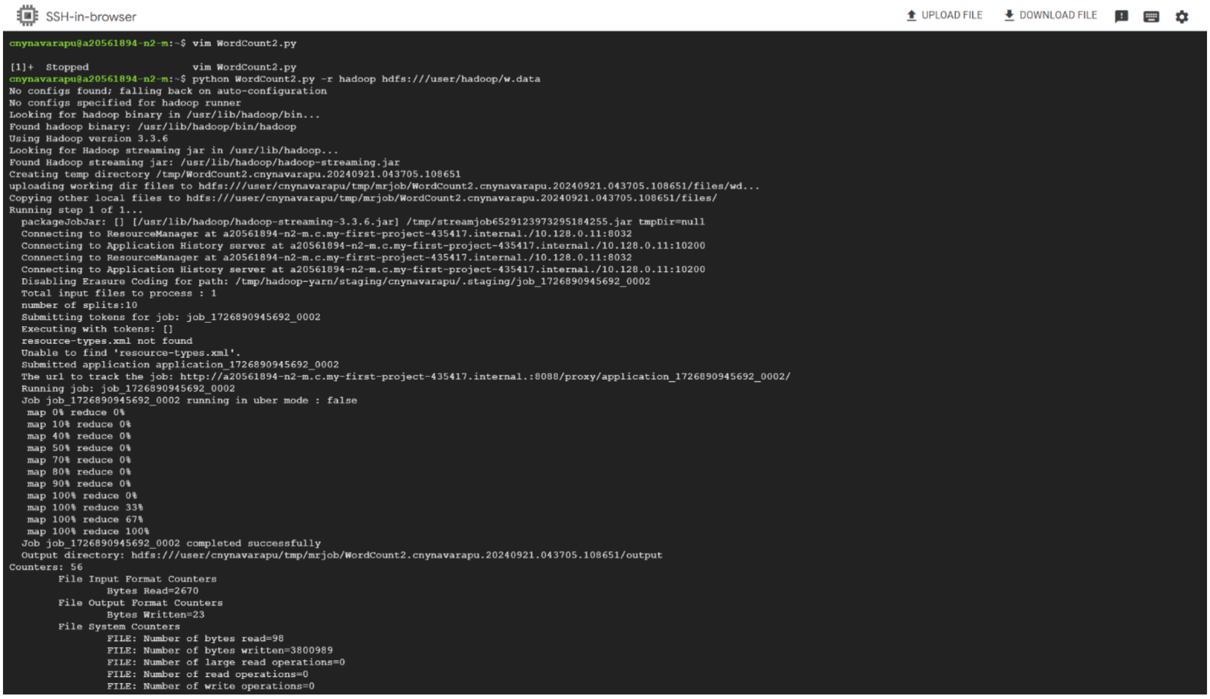


```
cnynavarapu@a20561894-n2-m:~$ python WordCount.py -r hadoop hdfs:///user/hadoop/w.data | head -n 10
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in /usr/lib/hadoop/bin...
Found hadoop binary: /usr/lib/hadoop/bin/hadoop
Using Hadoop version 3.3.6
Looking for Hadoop streaming jar in /usr/lib/hadoop...
Found Hadoop streaming jar: /usr/lib/hadoop/hadoop-streaming.jar
Creating temp directory /tmp/WordCount.cnynavarapu.20240921.042934.630679
uploading working dir files to hdfs:///user/cnynavarapu/tmp/mrjob/WordCount.cnynavarapu.20240921.042934.630679/files/wd...
Copying other local files to hdfs:///user/cnynavarapu/tmp/mrjob/WordCount.cnynavarapu.20240921.042934.630679/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.6.jar] /tmp/streamjob4979431168376685143.jar tmpDir=null
  Connecting to ResourceManager at a20561894-n2-m.c.my-first-project-435417.internal./10.128.0.11:8032
  Connecting to Application History server at a20561894-n2-m.c.my-first-project-435417.internal./10.128.0.11:10200
  Connecting to ResourceManager at a20561894-n2-m.c.my-first-project-435417.internal./10.128.0.11:8032
  Connecting to Application History server at a20561894-n2-m.c.my-first-project-435417.internal./10.128.0.11:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/cnynavarapu/.staging/job_1726890945692_0001
  Total input files to process : 1
  number of splits:10
  Submitting tokens for job: job_1726890945692_0001
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1726890945692_0001
  The url to track the job: http://a20561894-n2-m.c.my-first-project-435417.internal.:8088/proxy/application_1726890945692_0001/
  Running job: job_1726890945692_0001
  Job job_1726890945692_0001 running in uber mode : false
   map 0% reduce 0%
   map 10% reduce 0%
   map 20% reduce 0%
   map 30% reduce 0%
   map 40% reduce 0%
   map 50% reduce 0%
   map 60% reduce 0%
   map 70% reduce 0%
   map 80% reduce 0%
   map 100% reduce 0%
   map 100% reduce 33%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1726890945692_0001 completed successfully
  Output directory: hdfs:///user/cnynavarapu/tmp/mrjob/WordCount.cnynavarapu.20240921.042934.630679/output
Counters: 56
        File Input Format Counters
                Bytes Read=2670
        File Output Format Counters
                Bytes Written=652
        File System Counters
                FILE: Number of bytes read=954
                FILE: Number of bytes written=3802506
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3590
```

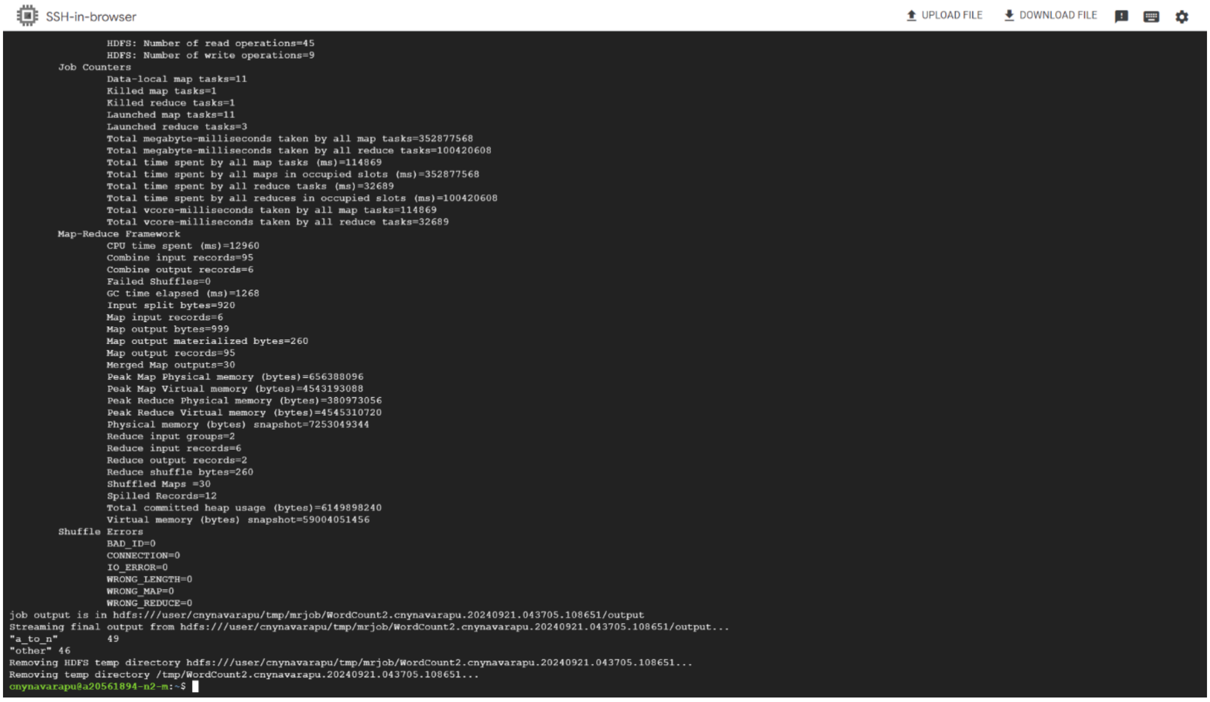


```
                Launched reduce tasks=3
                Total megabyte-milliseconds taken by all map tasks=365770752
                Total megabyte-milliseconds taken by all reduce tasks=102435840
                Total time spent by all map tasks (ms)=119066
                Total time spent by all maps in occupied slots (ms)=365770752
                Total time spent by all reduce tasks (ms)=33345
                Total time spent by all reduces in occupied slots (ms)=102435840
                Total vcore-milliseconds taken by all map tasks=119066
                Total vcore-milliseconds taken by all reduce tasks=33345
        Map-Reduce Framework
                CPU time spent (ms)=14900
                Combine input records=95
                Combine output records=80
                Failed Shuffles=0
                GC time elapsed (ms)=1276
                Input split bytes=920
                Map input records=6
                Map output bytes=891
                Map output materialized bytes=1116
                Map output records=95
                Merged Map outputs=30
                Peak Map Physical memory (bytes)=644808704
                Peak Map Virtual memory (bytes)=4549025792
                Peak Reduce Physical memory (bytes)=392335360
                Peak Reduce Virtual memory (bytes)=4545744896
                Physical memory (bytes) snapshot=7028518912
                Reduce input groups=65
                Reduce input records=80
                Reduce output records=65
                Reduce shuffle bytes=1116
                Shuffled Maps =30
                Spilled Records=160
                Total committed heap usage (bytes)=5911871488
                Virtual memory (bytes) snapshot=59022831616
        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/cnynavarapu/tmp/mrjob/WordCount.cnynavarapu.20240921.042934.630679/output
Streaming final output from hdfs:///user/cnynavarapu/tmp/mrjob/WordCount.cnynavarapu.20240921.042934.630679/output...
"an"    1
"are"   1
"available"     1
"by"    1
"combine"       1
"defined"       1
"dependencies"  1
"for"   1
"hadoop"        1
"job"   4
Removing HDFS temp directory hdfs:///user/cnynavarapu/tmp/mrjob/WordCount.cnynavarapu.20240921.042934.630679...
Removing temp directory /tmp/WordCount.cnynavarapu.20240921.042934.630679...
```

- Introduced vim command and modified the code for desired output

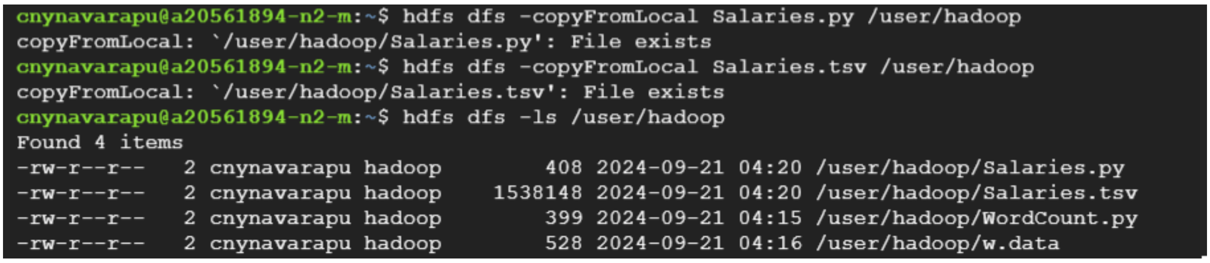


```
cnynavarapu@a20561894-n2-m:~$ vim WordCount2.py

[1]+  Stopped                 vim WordCount2.py
cnynavarapu@a20561894-n2-m:~$ 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 /usr/lib/hadoop/bin...
Found hadoop binary: /usr/lib/hadoop/bin/hadoop
Using Hadoop version 3.3.6
Looking for Hadoop streaming jar in /usr/lib/hadoop...
Found Hadoop streaming jar: /usr/lib/hadoop/hadoop-streaming.jar
Creating temp directory /tmp/WordCount2.cnynavarapu.20240921.043705.108651
uploading working dir files to hdfs:///user/cnynavarapu/tmp/mrjob/WordCount2.cnynavarapu.20240921.043705.108651/files/wd...
Copying other local files to hdfs:///user/cnynavarapu/tmp/mrjob/WordCount2.cnynavarapu.20240921.043705.108651/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.6.jar] /tmp/streamjob6529123973295184255.jar tmpDir=null
  Connecting to ResourceManager at a20561894-n2-m.c.my-first-project-435417.internal./10.128.0.11:8032
  Connecting to Application History server at a20561894-n2-m.c.my-first-project-435417.internal./10.128.0.11:10200
  Connecting to ResourceManager at a20561894-n2-m.c.my-first-project-435417.internal./10.128.0.11:8032
  Connecting to Application History server at a20561894-n2-m.c.my-first-project-435417.internal./10.128.0.11:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/cnynavarapu/.staging/job_1726890945692_0002
  Total input files to process : 1
  number of splits:10
  Submitting tokens for job: job_1726890945692_0002
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1726890945692_0002
  The url to track the job: http://a20561894-n2-m.c.my-first-project-435417.internal.:8088/proxy/application_1726890945692_0002/
  Running job: job_1726890945692_0002
  Job job_1726890945692_0002 running in uber mode : false
   map 0% reduce 0%
   map 10% reduce 0%
   map 40% reduce 0%
   map 50% reduce 0%
   map 70% reduce 0%
   map 80% reduce 0%
   map 90% reduce 0%
   map 100% reduce 0%
   map 100% reduce 33%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1726890945692_0002 completed successfully
  Output directory: hdfs:///user/cnynavarapu/tmp/mrjob/WordCount2.cnynavarapu.20240921.043705.108651/output
Counters: 56
        File Input Format Counters
                Bytes Read=2670
        File Output Format Counters
                Bytes Written=23
        File System Counters
                FILE: Number of bytes read=98
                FILE: Number of bytes written=3800989
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
```



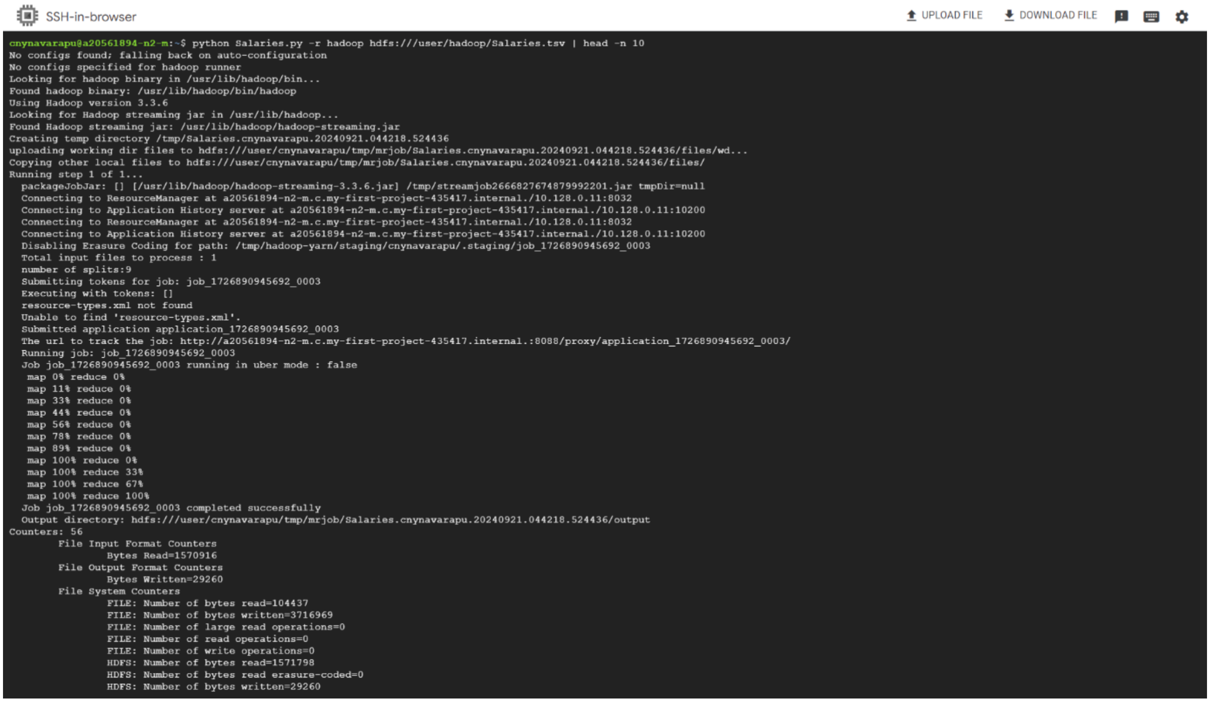
```
                HDFS: Number of read operations=45
                HDFS: Number of write operations=9
        Job Counters
                Data-local map tasks=11
                Killed map tasks=1
                Killed reduce tasks=1
                Launched map tasks=11
                Launched reduce tasks=3
                Total megabyte-milliseconds taken by all map tasks=352877568
                Total megabyte-milliseconds taken by all reduce tasks=100420608
                Total time spent by all map tasks (ms)=114869
                Total time spent by all maps in occupied slots (ms)=352877568
                Total time spent by all reduce tasks (ms)=32689
                Total time spent by all reduces in occupied slots (ms)=100420608
                Total vcore-milliseconds taken by all map tasks=114869
                Total vcore-milliseconds taken by all reduce tasks=32689
        Map-Reduce Framework
                CPU time spent (ms)=12960
                Combine input records=95
                Combine output records=6
                Failed Shuffles=0
                GC time elapsed (ms)=1268
                Input split bytes=920
                Map input records=6
                Map output bytes=999
                Map output materialized bytes=260
                Map output records=95
                Merged Map outputs=30
                Peak Map Physical memory (bytes)=656388096
                Peak Map Virtual memory (bytes)=4543193088
                Peak Reduce Physical memory (bytes)=380973056
                Peak Reduce Virtual memory (bytes)=4545310720
                Physical memory (bytes) snapshot=7253049344
                Reduce input groups=2
                Reduce input records=6
                Reduce output records=2
                Reduce shuffle bytes=260
                Shuffled Maps =30
                Spilled Records=12
                Total committed heap usage (bytes)=6149898240
                Virtual memory (bytes) snapshot=59004051456
        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/cnynavarapu/tmp/mrjob/WordCount2.cnynavarapu.20240921.043705.108651/output
Streaming final output from hdfs:///user/cnynavarapu/tmp/mrjob/WordCount2.cnynavarapu.20240921.043705.108651/output...
"a_to_n"        49
"other" 46
Removing HDFS temp directory hdfs:///user/cnynavarapu/tmp/mrjob/WordCount2.cnynavarapu.20240921.043705.108651...
Removing temp directory /tmp/WordCount2.cnynavarapu.20240921.043705.108651...
cnynavarapu@a20561894-n2-m:~$ 
```

- Uploaded Salaries.py and Salaries.tsv

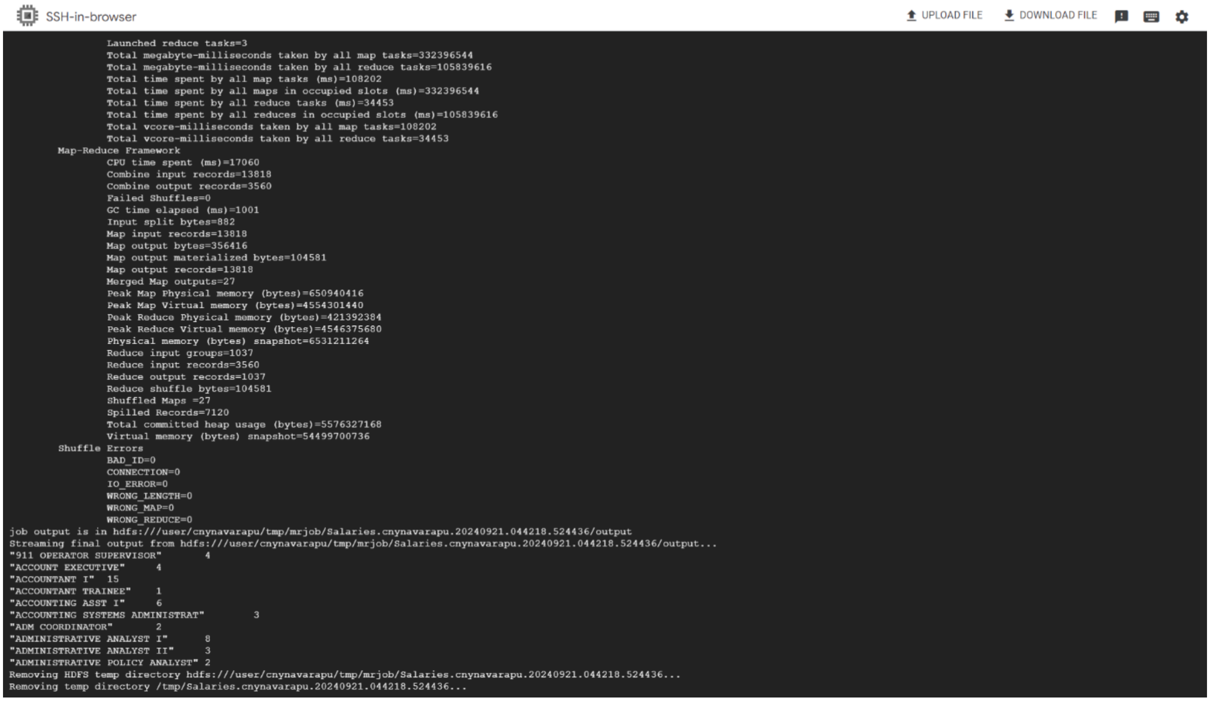


```
cnynavarapu@a20561894-n2-m:~$ hdfs dfs -copyFromLocal Salaries.py /user/hadoop
copyFromLocal: `/user/hadoop/Salaries.py': File exists
cnynavarapu@a20561894-n2-m:~$ hdfs dfs -copyFromLocal Salaries.tsv /user/hadoop
copyFromLocal: `/user/hadoop/Salaries.tsv': File exists
cnynavarapu@a20561894-n2-m:~$ hdfs dfs -ls /user/hadoop
Found 4 items
-rw-r--r--   2 cnynavarapu hadoop        408 2024-09-21 04:20 /user/hadoop/Salaries.py
-rw-r--r--   2 cnynavarapu hadoop    1538148 2024-09-21 04:20 /user/hadoop/Salaries.tsv
-rw-r--r--   2 cnynavarapu hadoop        399 2024-09-21 04:15 /user/hadoop/WordCount.py
-rw-r--r--   2 cnynavarapu hadoop        528 2024-09-21 04:16 /user/hadoop/w.data
```

- Tested that Salaries.py file and printed 10 values

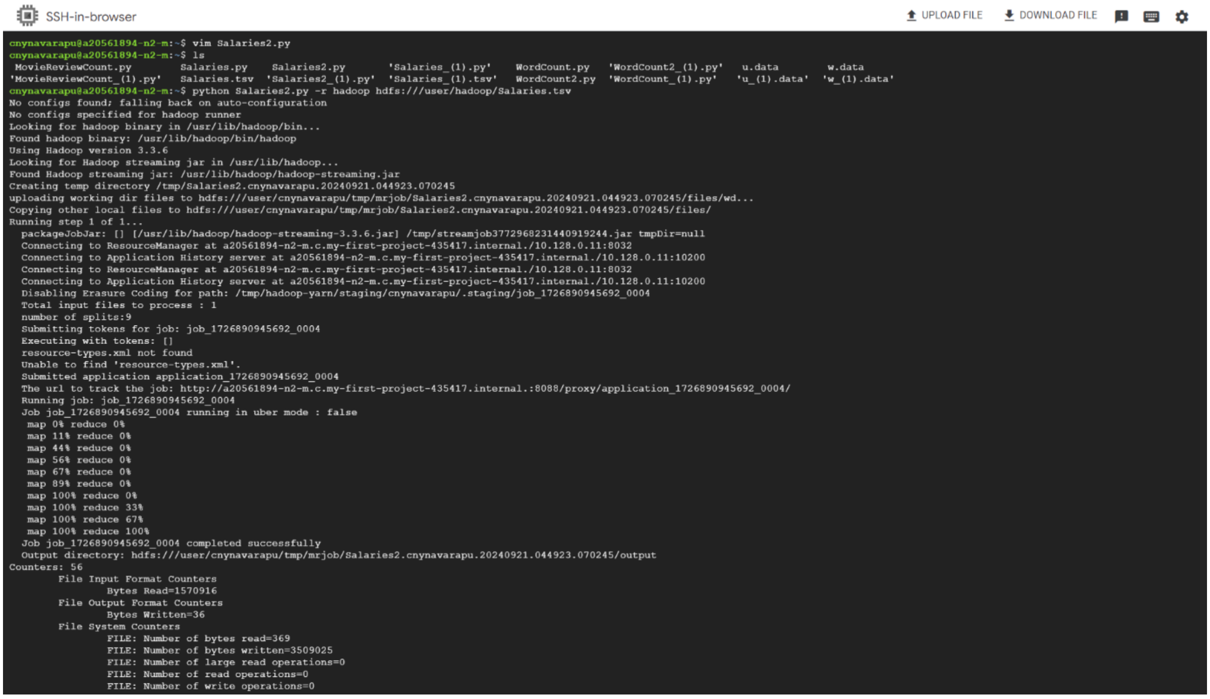


```
cnynavarapu@a20561894-n2-m:~$ python Salaries.py -r hadoop hdfs:///user/hadoop/Salaries.tsv | head -n 10
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in /usr/lib/hadoop/bin...
Found hadoop binary: /usr/lib/hadoop/bin/hadoop
Using Hadoop version 3.3.6
Looking for Hadoop streaming jar in /usr/lib/hadoop...
Found Hadoop streaming jar: /usr/lib/hadoop/hadoop-streaming.jar
Creating temp directory /tmp/Salaries.cnynavarapu.20240921.044218.524436
uploading working dir files to hdfs:///user/cnynavarapu/tmp/mrjob/Salaries.cnynavarapu.20240921.044218.524436/files/wd...
Copying other local files to hdfs:///user/cnynavarapu/tmp/mrjob/Salaries.cnynavarapu.20240921.044218.524436/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.6.jar] /tmp/streamjob2666827674879992201.jar tmpDir=null
  Connecting to ResourceManager at a20561894-n2-m.c.my-first-project-435417.internal./10.128.0.11:8032
  Connecting to Application History server at a20561894-n2-m.c.my-first-project-435417.internal./10.128.0.11:10200
  Connecting to ResourceManager at a20561894-n2-m.c.my-first-project-435417.internal./10.128.0.11:8032
  Connecting to Application History server at a20561894-n2-m.c.my-first-project-435417.internal./10.128.0.11:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/cnynavarapu/.staging/job_1726890945692_0003
  Total input files to process : 1
  number of splits:9
  Submitting tokens for job: job_1726890945692_0003
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1726890945692_0003
  The url to track the job: http://a20561894-n2-m.c.my-first-project-435417.internal.:8088/proxy/application_1726890945692_0003/
  Running job: job_1726890945692_0003
  Job job_1726890945692_0003 running in uber mode : false
   map 0% reduce 0%
   map 11% reduce 0%
   map 33% reduce 0%
   map 44% reduce 0%
   map 56% reduce 0%
   map 78% reduce 0%
   map 89% reduce 0%
   map 100% reduce 0%
   map 100% reduce 33%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1726890945692_0003 completed successfully
  Output directory: hdfs:///user/cnynavarapu/tmp/mrjob/Salaries.cnynavarapu.20240921.044218.524436/output
Counters: 56
        File Input Format Counters
                Bytes Read=1570916
        File Output Format Counters
                Bytes Written=29260
        File System Counters
                FILE: Number of bytes read=104437
                FILE: Number of bytes written=3716969
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1571798
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=29260
```

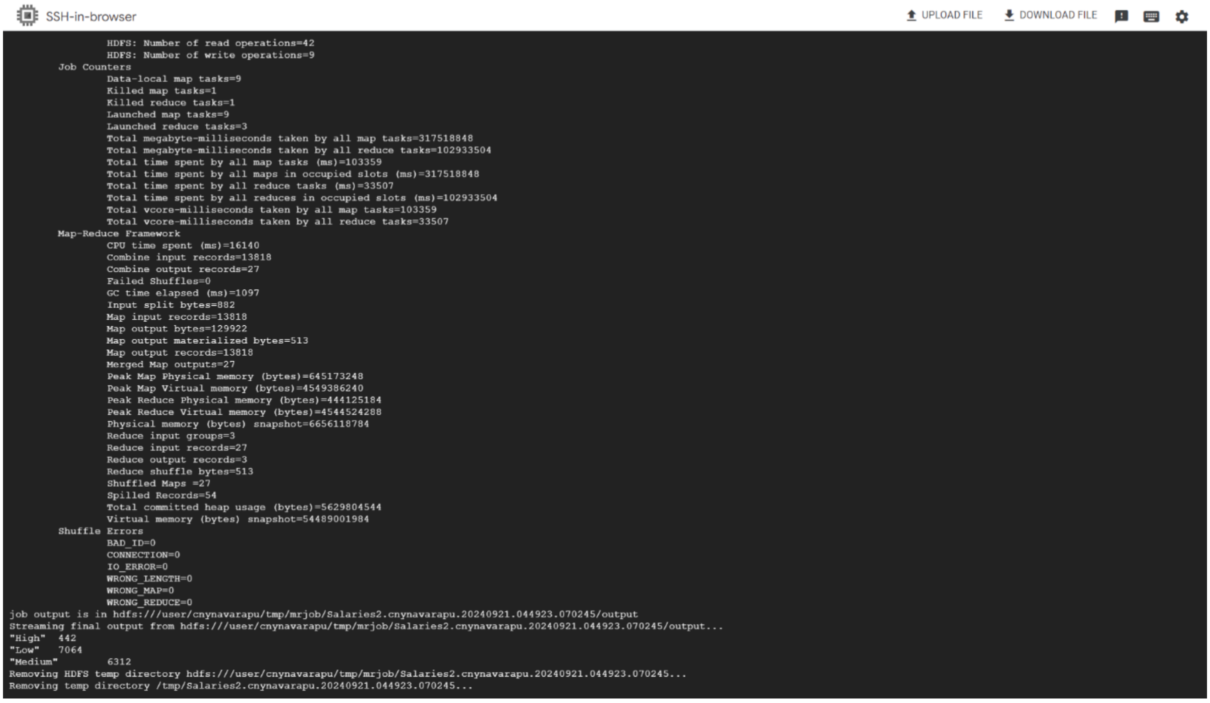


```
                Launched reduce tasks=3
                Total megabyte-milliseconds taken by all map tasks=332396544
                Total megabyte-milliseconds taken by all reduce tasks=105839616
                Total time spent by all map tasks (ms)=108202
                Total time spent by all maps in occupied slots (ms)=332396544
                Total time spent by all reduce tasks (ms)=34453
                Total time spent by all reduces in occupied slots (ms)=105839616
                Total vcore-milliseconds taken by all map tasks=108202
                Total vcore-milliseconds taken by all reduce tasks=34453
        Map-Reduce Framework
                CPU time spent (ms)=17060
                Combine input records=13818
                Combine output records=3560
                Failed Shuffles=0
                GC time elapsed (ms)=1001
                Input split bytes=882
                Map input records=13818
                Map output bytes=356416
                Map output materialized bytes=104581
                Map output records=13818
                Merged Map outputs=27
                Peak Map Physical memory (bytes)=650940416
                Peak Map Virtual memory (bytes)=4554301440
                Peak Reduce Physical memory (bytes)=421392384
                Peak Reduce Virtual memory (bytes)=4546375680
                Physical memory (bytes) snapshot=6531211264
                Reduce input groups=1037
                Reduce input records=3560
                Reduce output records=1037
                Reduce shuffle bytes=104581
                Shuffled Maps =27
                Spilled Records=7120
                Total committed heap usage (bytes)=5576327168
                Virtual memory (bytes) snapshot=54499700736
        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/cnynavarapu/tmp/mrjob/Salaries.cnynavarapu.20240921.044218.524436/output
Streaming final output from hdfs:///user/cnynavarapu/tmp/mrjob/Salaries.cnynavarapu.20240921.044218.524436/output...
"911 OPERATOR SUPERVISOR"       4
"ACCOUNT EXECUTIVE"     4
"ACCOUNTANT I"  15
"ACCOUNTANT TRAINEE"    1
"ACCOUNTING ASST I"     6
"ACCOUNTING SYSTEMS ADMINISTRAT"        3
"ADM COORDINATOR"       2
"ADMINISTRATIVE ANALYST I"      8
"ADMINISTRATIVE ANALYST II"     3
"ADMINISTRATIVE POLICY ANALYST" 2
Removing HDFS temp directory hdfs:///user/cnynavarapu/tmp/mrjob/Salaries.cnynavarapu.20240921.044218.524436...
Removing temp directory /tmp/Salaries.cnynavarapu.20240921.044218.524436...
```

- Now edited the code using vim and named it as Salaries2.py

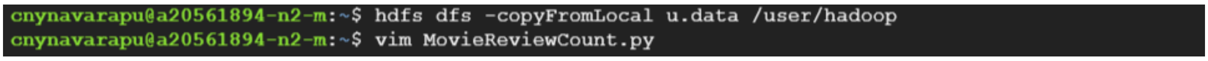


```
cnynavarapu@a20561894-n2-m:~$ vim Salaries2.py
cnynavarapu@a20561894-n2-m:~$ ls
 MovieReviewCount.py        Salaries.py    Salaries2.py       'Salaries_(1).py'    WordCount.py   'WordCount2_(1).py'   u.data        w.data
'MovieReviewCount_(1).py'   Salaries.tsv  'Salaries2_(1).py'  'Salaries_(1).tsv'   WordCount2.py  'WordCount_(1).py'   'u_(1).data'  'w_(1).data'
cnynavarapu@a20561894-n2-m:~$ 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 /usr/lib/hadoop/bin...
Found hadoop binary: /usr/lib/hadoop/bin/hadoop
Using Hadoop version 3.3.6
Looking for Hadoop streaming jar in /usr/lib/hadoop...
Found Hadoop streaming jar: /usr/lib/hadoop/hadoop-streaming.jar
Creating temp directory /tmp/Salaries2.cnynavarapu.20240921.044923.070245
uploading working dir files to hdfs:///user/cnynavarapu/tmp/mrjob/Salaries2.cnynavarapu.20240921.044923.070245/files/wd...
Copying other local files to hdfs:///user/cnynavarapu/tmp/mrjob/Salaries2.cnynavarapu.20240921.044923.070245/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.6.jar] /tmp/streamjob3772968231440919244.jar tmpDir=null
  Connecting to ResourceManager at a20561894-n2-m.c.my-first-project-435417.internal./10.128.0.11:8032
  Connecting to Application History server at a20561894-n2-m.c.my-first-project-435417.internal./10.128.0.11:10200
  Connecting to ResourceManager at a20561894-n2-m.c.my-first-project-435417.internal./10.128.0.11:8032
  Connecting to Application History server at a20561894-n2-m.c.my-first-project-435417.internal./10.128.0.11:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/cnynavarapu/.staging/job_1726890945692_0004
  Total input files to process : 1
  number of splits:9
  Submitting tokens for job: job_1726890945692_0004
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1726890945692_0004
  The url to track the job: http://a20561894-n2-m.c.my-first-project-435417.internal.:8088/proxy/application_1726890945692_0004/
  Running job: job_1726890945692_0004
  Job job_1726890945692_0004 running in uber mode : false
   map 0% reduce 0%
   map 11% reduce 0%
   map 44% reduce 0%
   map 56% reduce 0%
   map 67% reduce 0%
   map 89% reduce 0%
   map 100% reduce 0%
   map 100% reduce 33%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1726890945692_0004 completed successfully
  Output directory: hdfs:///user/cnynavarapu/tmp/mrjob/Salaries2.cnynavarapu.20240921.044923.070245/output
Counters: 56
        File Input Format Counters
                Bytes Read=1570916
        File Output Format Counters
                Bytes Written=36
        File System Counters
                FILE: Number of bytes read=369
                FILE: Number of bytes written=3509025
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
```

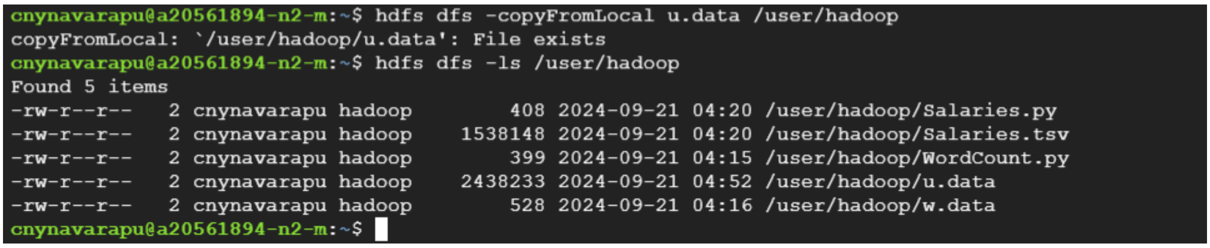


```
                HDFS: Number of read operations=42
                HDFS: Number of write operations=9
        Job Counters
                Data-local map tasks=9
                Killed map tasks=1
                Killed reduce tasks=1
                Launched map tasks=9
                Launched reduce tasks=3
                Total megabyte-milliseconds taken by all map tasks=317518848
                Total megabyte-milliseconds taken by all reduce tasks=102933504
                Total time spent by all map tasks (ms)=103359
                Total time spent by all maps in occupied slots (ms)=317518848
                Total time spent by all reduce tasks (ms)=33507
                Total time spent by all reduces in occupied slots (ms)=102933504
                Total vcore-milliseconds taken by all map tasks=103359
                Total vcore-milliseconds taken by all reduce tasks=33507
        Map-Reduce Framework
                CPU time spent (ms)=16140
                Combine input records=13818
                Combine output records=27
                Failed Shuffles=0
                GC time elapsed (ms)=1097
                Input split bytes=882
                Map input records=13818
                Map output bytes=129922
                Map output materialized bytes=513
                Map output records=13818
                Merged Map outputs=27
                Peak Map Physical memory (bytes)=645173248
                Peak Map Virtual memory (bytes)=4549386240
                Peak Reduce Physical memory (bytes)=444125184
                Peak Reduce Virtual memory (bytes)=4544524288
                Physical memory (bytes) snapshot=6656118784
                Reduce input groups=3
                Reduce input records=27
                Reduce output records=3
                Reduce shuffle bytes=513
                Shuffled Maps =27
                Spilled Records=54
                Total committed heap usage (bytes)=5629804544
                Virtual memory (bytes) snapshot=54489001984
        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/cnynavarapu/tmp/mrjob/Salaries2.cnynavarapu.20240921.044923.070245/output
Streaming final output from hdfs:///user/cnynavarapu/tmp/mrjob/Salaries2.cnynavarapu.20240921.044923.070245/output...
"High"  442
"Low"   7064
"Medium"        6312
Removing HDFS temp directory hdfs:///user/cnynavarapu/tmp/mrjob/Salaries2.cnynavarapu.20240921.044923.070245...
Removing temp directory /tmp/Salaries2.cnynavarapu.20240921.044923.070245...
```

3. Loaded u.data file

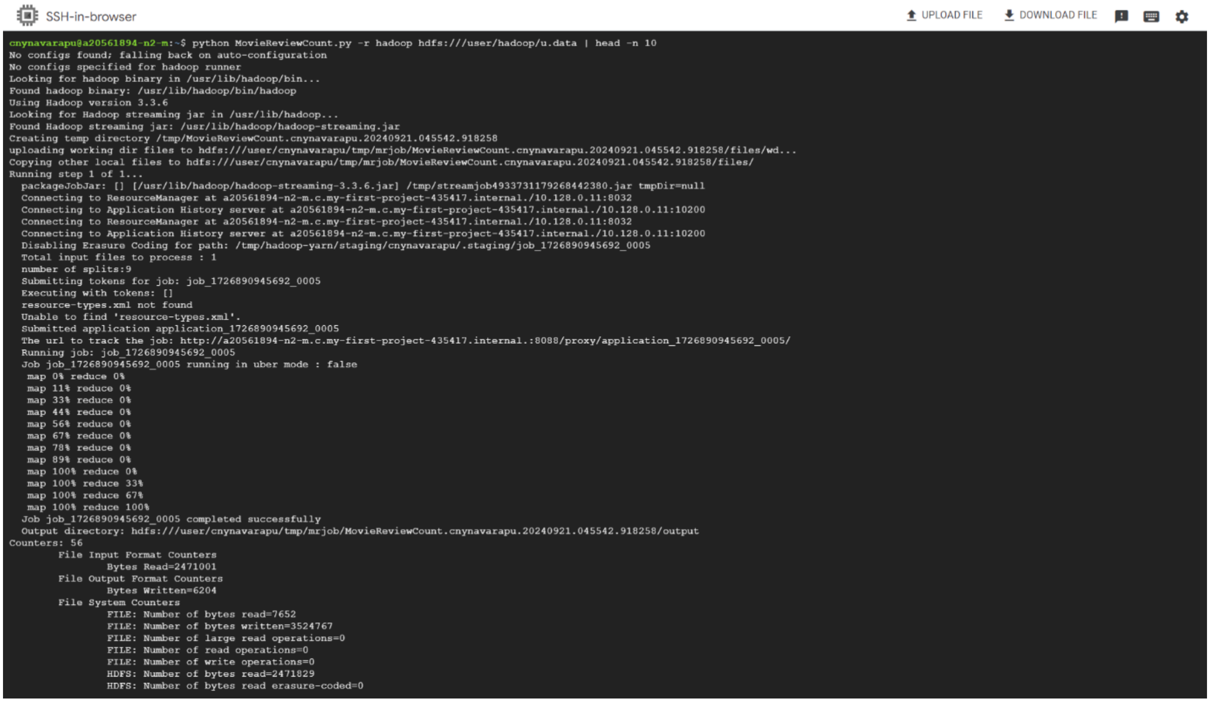


```
cnynavarapu@a20561894-n2-m:~$ hdfs dfs -copyFromLocal u.data /user/hadoop
cnynavarapu@a20561894-n2-m:~$ vim MovieReviewCount.py
```

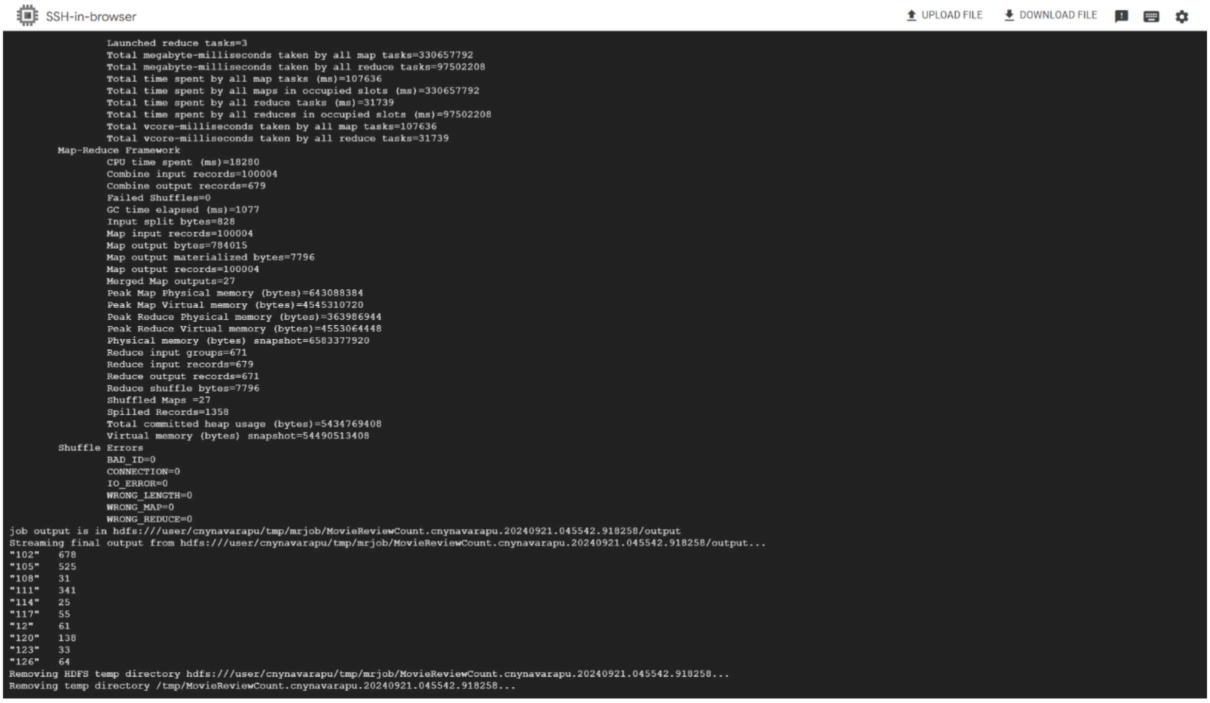


```
cnynavarapu@a20561894-n2-m:~$ hdfs dfs -copyFromLocal u.data /user/hadoop
copyFromLocal: `/user/hadoop/u.data': File exists
cnynavarapu@a20561894-n2-m:~$ hdfs dfs -ls /user/hadoop
Found 5 items
-rw-r--r--   2 cnynavarapu hadoop        408 2024-09-21 04:20 /user/hadoop/Salaries.py
-rw-r--r--   2 cnynavarapu hadoop    1538148 2024-09-21 04:20 /user/hadoop/Salaries.tsv
-rw-r--r--   2 cnynavarapu hadoop        399 2024-09-21 04:15 /user/hadoop/WordCount.py
-rw-r--r--   2 cnynavarapu hadoop    2438233 2024-09-21 04:52 /user/hadoop/u.data
-rw-r--r--   2 cnynavarapu hadoop        528 2024-09-21 04:16 /user/hadoop/w.data
cnynavarapu@a20561894-n2-m:~$ 
```

4. Written the required code in VIM to get the desired results output



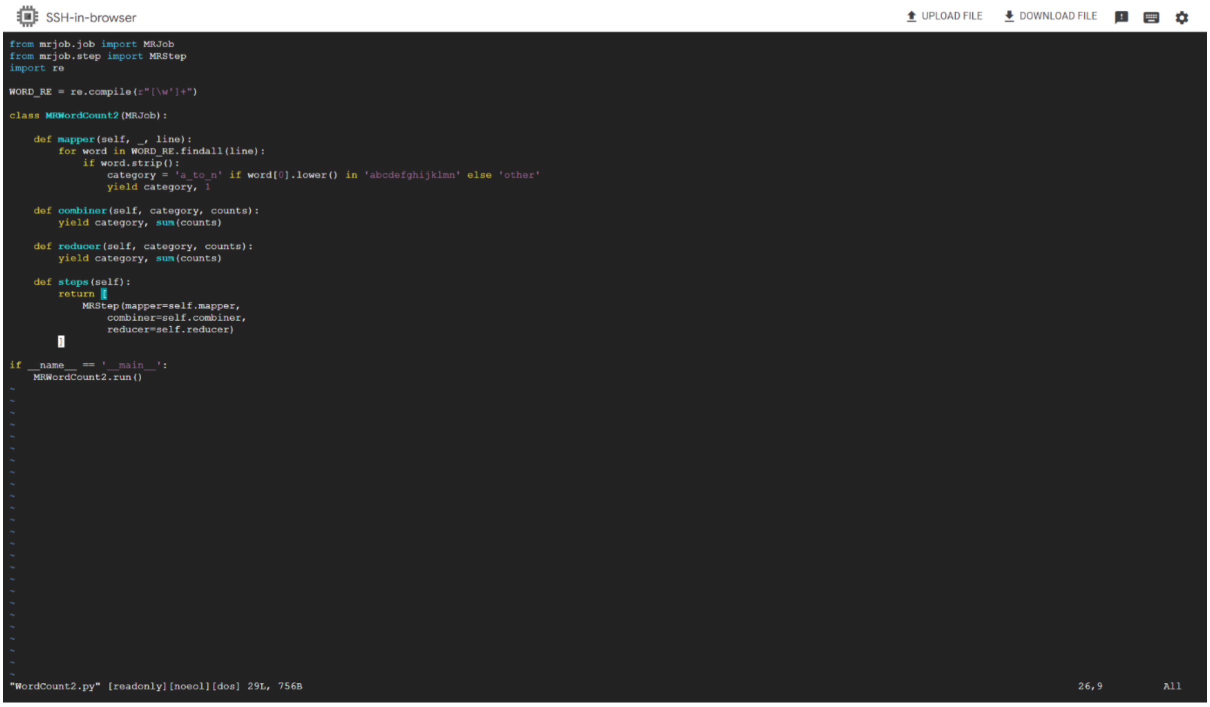
```
cnynavarapu@a20561894-n2-m:~$ python MovieReviewCount.py -r hadoop hdfs:///user/hadoop/u.data | head -n 10
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in /usr/lib/hadoop/bin...
Found hadoop binary: /usr/lib/hadoop/bin/hadoop
Using Hadoop version 3.3.6
Looking for Hadoop streaming jar in /usr/lib/hadoop...
Found Hadoop streaming jar: /usr/lib/hadoop/hadoop-streaming.jar
Creating temp directory /tmp/MovieReviewCount.cnynavarapu.20240921.045542.918258
uploading working dir files to hdfs:///user/cnynavarapu/tmp/mrjob/MovieReviewCount.cnynavarapu.20240921.045542.918258/files/wd...
Copying other local files to hdfs:///user/cnynavarapu/tmp/mrjob/MovieReviewCount.cnynavarapu.20240921.045542.918258/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.6.jar] /tmp/streamjob4933731179268442380.jar tmpDir=null
  Connecting to ResourceManager at a20561894-n2-m.c.my-first-project-435417.internal./10.128.0.11:8032
  Connecting to Application History server at a20561894-n2-m.c.my-first-project-435417.internal./10.128.0.11:10200
  Connecting to ResourceManager at a20561894-n2-m.c.my-first-project-435417.internal./10.128.0.11:8032
  Connecting to Application History server at a20561894-n2-m.c.my-first-project-435417.internal./10.128.0.11:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/cnynavarapu/.staging/job_1726890945692_0005
  Total input files to process : 1
  number of splits:9
  Submitting tokens for job: job_1726890945692_0005
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1726890945692_0005
  The url to track the job: http://a20561894-n2-m.c.my-first-project-435417.internal.:8088/proxy/application_1726890945692_0005/
  Running job: job_1726890945692_0005
  Job job_1726890945692_0005 running in uber mode : false
   map 0% reduce 0%
   map 11% reduce 0%
   map 33% reduce 0%
   map 44% reduce 0%
   map 56% reduce 0%
   map 67% reduce 0%
   map 78% reduce 0%
   map 89% reduce 0%
   map 100% reduce 0%
   map 100% reduce 33%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1726890945692_0005 completed successfully
  Output directory: hdfs:///user/cnynavarapu/tmp/mrjob/MovieReviewCount.cnynavarapu.20240921.045542.918258/output
Counters: 56
        File Input Format Counters
                Bytes Read=2471001
        File Output Format Counters
                Bytes Written=6204
        File System Counters
                FILE: Number of bytes read=7652
                FILE: Number of bytes written=3524767
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2471829
                HDFS: Number of bytes read erasure-coded=0
```



```
                Launched reduce tasks=3
                Total megabyte-milliseconds taken by all map tasks=330657792
                Total megabyte-milliseconds taken by all reduce tasks=97502208
                Total time spent by all map tasks (ms)=107636
                Total time spent by all maps in occupied slots (ms)=330657792
                Total time spent by all reduce tasks (ms)=31739
                Total time spent by all reduces in occupied slots (ms)=97502208
                Total vcore-milliseconds taken by all map tasks=107636
                Total vcore-milliseconds taken by all reduce tasks=31739
        Map-Reduce Framework
                CPU time spent (ms)=18280
                Combine input records=100004
                Combine output records=679
                Failed Shuffles=0
                GC time elapsed (ms)=1077
                Input split bytes=828
                Map input records=100004
                Map output bytes=784015
                Map output materialized bytes=7796
                Map output records=100004
                Merged Map outputs=27
                Peak Map Physical memory (bytes)=643088384
                Peak Map Virtual memory (bytes)=4545310720
                Peak Reduce Physical memory (bytes)=363986944
                Peak Reduce Virtual memory (bytes)=4553064448
                Physical memory (bytes) snapshot=6583377920
                Reduce input groups=671
                Reduce input records=679
                Reduce output records=671
                Reduce shuffle bytes=7796
                Shuffled Maps =27
                Spilled Records=1358
                Total committed heap usage (bytes)=5434769408
                Virtual memory (bytes) snapshot=54490513408
        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/cnynavarapu/tmp/mrjob/MovieReviewCount.cnynavarapu.20240921.045542.918258/output
Streaming final output from hdfs:///user/cnynavarapu/tmp/mrjob/MovieReviewCount.cnynavarapu.20240921.045542.918258/output...
"102"   678
"105"   525
"108"   31
"111"   341
"114"   25
"117"   55
"12"    61
"120"   138
"123"   33
"126"   64
Removing HDFS temp directory hdfs:///user/cnynavarapu/tmp/mrjob/MovieReviewCount.cnynavarapu.20240921.045542.918258...
Removing temp directory /tmp/MovieReviewCount.cnynavarapu.20240921.045542.918258...
```

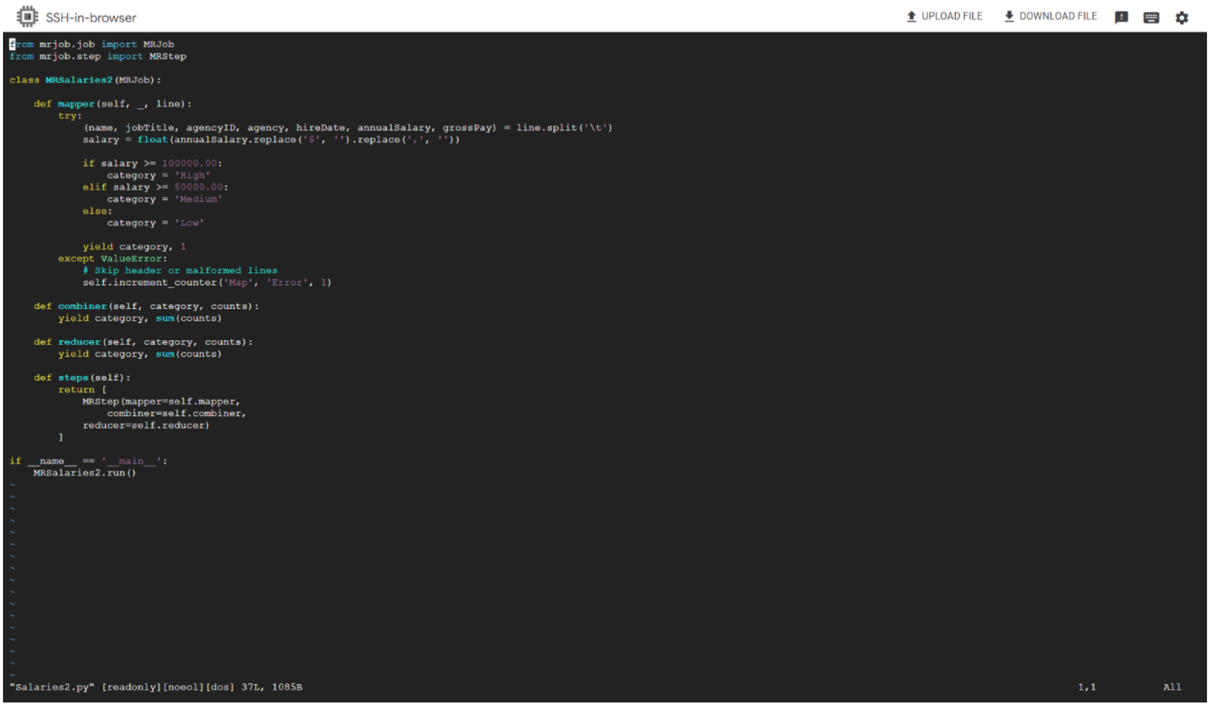
Code Screenshots which are used:

- WordCount2.py



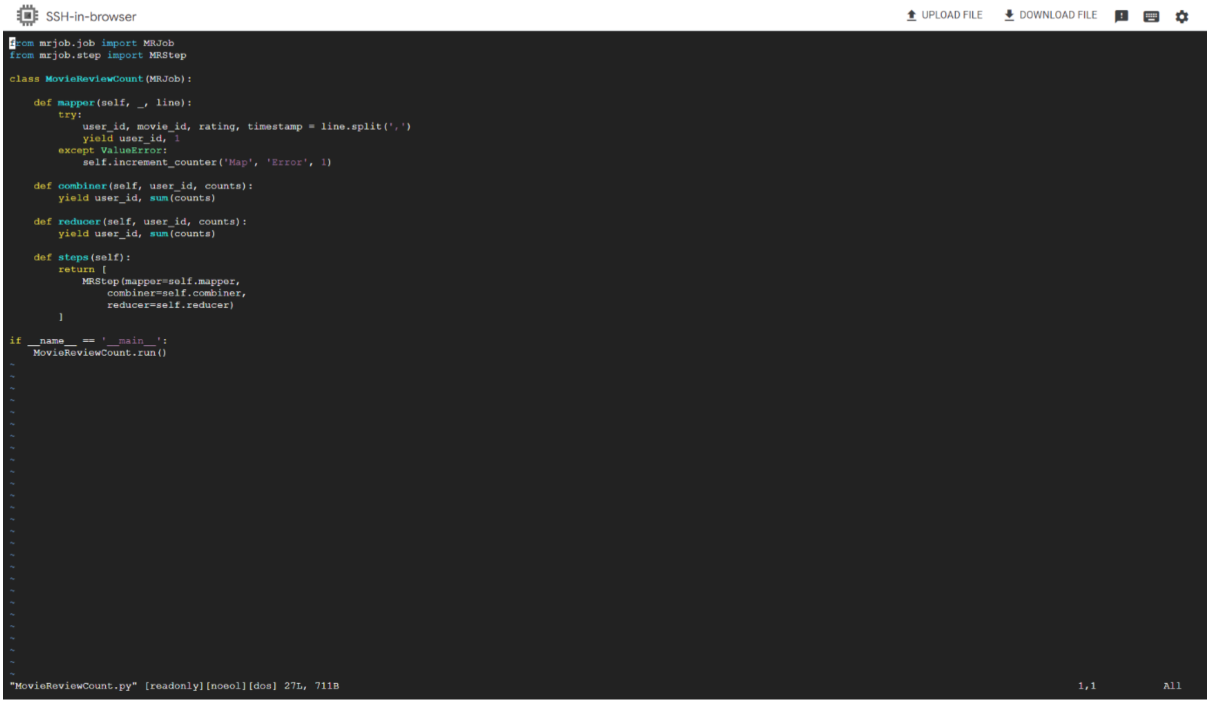
```
from mrjob.job import MRJob
from mrjob.step import MRStep
import re

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

class MRWordCount2(MRJob):

    def mapper(self, _, line):
        for word in WORD_RE.findall(line):
            if word.strip():
                category = 'a_to_n' if word[0].lower() in 'abcdefghijklmn' else 'other'
                yield category, 1

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

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

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

if __name__ == '__main__':
    MRWordCount2.run()

"WordCount2.py" [readonly][noeol][dos] 29L, 756B                                    26,9          All
```

- Salaries2.py



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

class MRSalaries2(MRJob):

    def mapper(self, _, line):
        try:
            (name, jobTitle, agencyID, agency, hireDate, annualSalary, grossPay) = line.split('\t')
            salary = float(annualSalary.replace('$', '').replace(',', ''))

            if salary >= 100000.00:
                category = 'High'
            elif salary >= 50000.00:
                category = 'Medium'
            else:
                category = 'Low'

            yield category, 1
        except ValueError:
            # Skip header or malformed lines
            self.increment_counter('Map', 'Error', 1)

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

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

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

if __name__ == '__main__':
    MRSalaries2.run()

"Salaries2.py" [readonly][noeol][dos] 37L, 1085B                                    1,1           All
```

- MovieReviewCount.py



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

class MovieReviewCount(MRJob):

    def mapper(self, _, line):
        try:
            user_id, movie_id, rating, timestamp = line.split(',')
            yield user_id, 1
        except ValueError:
            self.increment_counter('Map', 'Error', 1)

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

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

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

if __name__ == '__main__':
    MovieReviewCount.run()
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
"MovieReviewCount.py" [readonly][noeol][dos] 27L, 711B                    1,1           All
```