

# Big Data Technologies CSP-554

## Project Report Document

Vaishnavi Prasanna Shetty  
A20519894

[vprasannashetty@hawk.iit.edu](mailto:vprasannashetty@hawk.iit.edu)

<https://github.com/VFA22SCM94P/CSP554>

### **Project Name: Bigdata File Filter and Convertor**

#### **Abstract:**

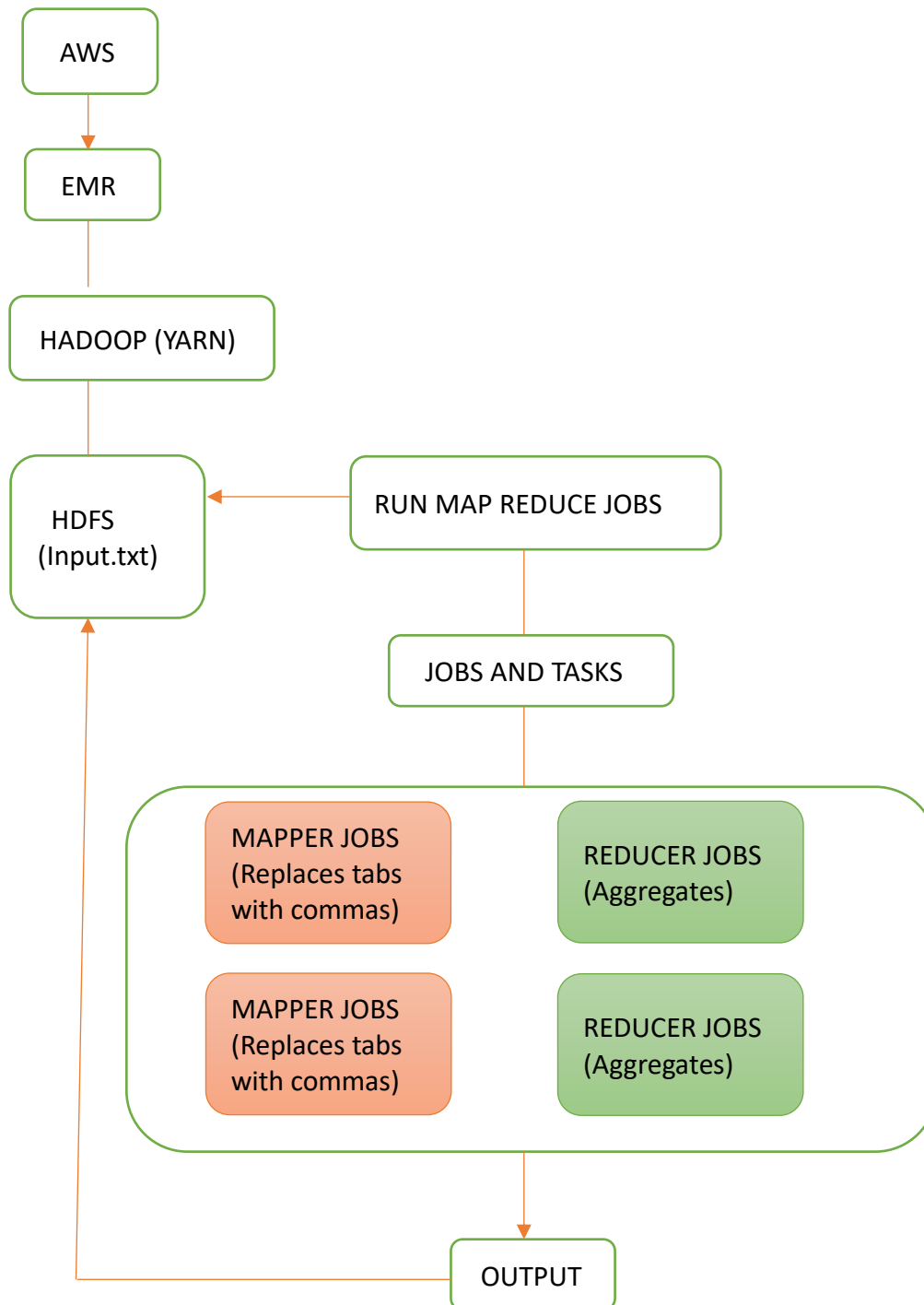
The work involves contribution towards the existing bigdata-file-viewer project by Eugene-Mark which is a cross platform desktop application to view the common bigdata binary format like Parquet, Avro.

The main objective of this project was to develop a program that converts the file format from Comma Separated Value (CSV) to Tab Separated Value (TSV) format. Map Reduce concept is utilized to provide the deliverables.

#### **Overview:**

The program has Mapper and Reducer jobs which gets the file from HDFS and modifies the file format from Comma Separated Values to Tab Separated Values. The file is fetched from HDFS location. The program mainly checks three important things before the mapper job is run which is confirming whether the given file is not empty, making sure that the file consists of commas prior running mapper and reducer jobs to avoid map reduce job runs.

## Architecture:



## **Design:**

There are 4 major components:

- Hadoop Distributed File System (HDFS): The input file is placed in HDFS.
- AWS Elastic MapReduce (EMR): This is used for creating the cluster for running the MapReduce jobs.
- MapReduce Job: This replaces the commas with the tabs for the input file.

MapReduce Job:

Mapper:

The mapper reads the input data from HDFS and processes it concurrently throughout the cluster during the map phase. The input data for the mapper is provided in the form of key-value pairs, where the key is the record's input file byte offset and the value is the actual record. Each entry is processed by the mapper by substituting tab spaces for commas using the `replace()` method.

Reducer:

The output from the mapper is delivered to the reducer, which processes it concurrently across the cluster during the reduce phase. Key-value pairs are provided to the reducer as input data, with the key being the dummy value produced by the mapper and the value being the updated record. The output of the reducer is written to HDFS as a tab-separated file after combining all the updated records together.

## Code:

### mapper.py

```
mapper.py — hadoop [SSH: ec2-54-210-197-101.compute-1.amazonaws.com]
mapper.py • reducer.py NoCommaFile.txt part-00000 •
mapper.py
1  import os
2  import sys
3  from mrjob.job import MRJob
4
5  class MRReplaceComma (MRJob):
6      def mapper(self, _, line):
7          yield None, line.strip().replace(',', ' ')
8
9  if __name__ == '__main__':
10
11      filename = 'data.csv'
12      file_size = os.path.getsize(filename)
13
14      file_size_mb = file_size / (1024 * 1024)
15
16      # if file_size == 0:
17
18      if os.path.getsize(filename) == 0:
19          print("File is empty!")
20          sys.exit()
21
22      # program to check input file size .
23      # because the EB volume on EMR cluster has a limit.
24      # so this program tests if the file is under 400MB
25
26      if file_size_mb > 400:
27          print(f"The file size greater than EBS volume .")
28          # exit from execution if the file size exceeds
29          sys.exit()
30      else:
31          # push to HDFS();
32          print(f"The File size acceptable, Pushed to HDFS.")
33
34      try:
35          with open(filename, 'r') as f:
36              # read the contents of the file
37              contents = f.read()
38              count = sum(line.count(",") for line in f)
39              if count == 0 :
40                  print("File does not have any commas to replace with space")
41
42      except FileNotFoundError:
43          # handle the file not found exception here
44          print("FileNotFoundError.")
45
46      MRReplaceComma.run()
```

### reducer.py

```
reducer.py > ...
1  from mrjob.job import MRJob
2  class MRReplaceComma(MRJob):
3      def reducer(self, _, values):
4          for value in values:
5              yield None, value.strip()
6
7  if __name__ == '__main__':
8      MRReplaceComma.run()
```

## File generation code

```
filegen.py > ...
1  import csv
2  import random
3
4  with open('data.csv', 'w') as csvfile:
5      writer = csv.writer(csvfile)
6      for i in range(100000):
7          row = [random.random() for j in range(10)]
8          writer.writerow(row)
9
```

Code to generate file without any comma in it.

```
nocomma.py > ...
1  import random
2
3  filename = 'NoCommaFile.txt'
4
5  # data without commas
6  data = ''.join(random.choices('12345', k=10000))
7
8  # Write the data to the file
9  with open(filename, 'w') as f:
10     f.write(data)
11
12  print(f"The file created")
13
```

### Functioning of the code on the small data sample

Input file

```
≡ simpleinput.txt ×
≡ simpleinput.txt
1      A,B,C,D
2      A,B,C,D
3      A,B,C,D
4      A,B,C,D
5      A,B,C,D
6      A,B,C,D
7      A,B,C,D
8      A,B,C,D
9      A,B,C,D
10     A,B,C,D
11     A,B,C,D
12     A,B,C,D
13     A,B,C,D
14     A,B,C,D
15     A,B,C,D
```

Output:

[illegible]

## Job run screenshots:

Runtime: 38s

```
● [hadoop@ip-172-31-31-170 ~]$ time python mapper.py -r hadoop hdfs:///user/hadoop/simpleinput.txt --output-dir hdfs:///outputs --jobconf mapreduce.job.reduces=1 && pyt
hon reducer.py -r hadoop hdfs:///outputs/* --output-dir hdfs:///final_outputs
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/mapper.hadoop.20230430.222945.327419
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/mapper.hadoop.20230430.222945.327419/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/mapper.hadoop.20230430.222945.327419/files/
Running step 1 of 1...
packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-4.jar] /tmp/streamjob1794048961707718496.jar tmpDir=null
Connecting to ResourceManager at ip-172-31-31-170.ec2.internal/172.31.31.170:8032
Connecting to Application History server at ip-172-31-31-170.ec2.internal/172.31.31.170:10200
Connecting to ResourceManager at ip-172-31-31-170.ec2.internal/172.31.31.170:8032
Connecting to Application History server at ip-172-31-31-170.ec2.internal/172.31.31.170:10200
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_1682892957887_0002
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_1682892957887_0002
The url to track the job: http://ip-172-31-31-170.ec2.internal:20888/proxy/application_1682892957887_0002/
Running job: job_1682892957887_0002
Job job_1682892957887_0002 running in uber mode : false
map 0% reduce 0%
map 50% reduce 0%
map 100% reduce 0%
map 100% reduce 100%
Job job_1682892957887_0002 completed successfully
Output directory: hdfs:///outputs
Counters: 50
```

```
job output is in hdfs:///outputs
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/mapper.hadoop.20230430.222945.327419...
Removing temp directory /tmp/mapper.hadoop.20230430.222945.327419...

real    1m15.843s
user    0m38.298s
sys     0m2.643s
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/reducer.hadoop.20230430.223102.103412
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/reducer.hadoop.20230430.223102.103412/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/reducer.hadoop.20230430.223102.103412/files/
Running step 1 of 1...
```

```
Map output bytes=375
Map output materialized bytes=193
Map output records=15
Merged Map outputs=4
Physical memory (bytes) snapshot=2032308224
Reduce input groups=1
Reduce input records=15
Reduce output records=15
Reduce shuffle bytes=193
Shuffled Maps =4
Spilled Records=30
Total committed heap usage (bytes)=1560805376
Virtual memory (bytes) snapshot=17867657216
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
job output is in hdfs:///final_outputs
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/reducer.hadoop.20230430.223102.103412...
Removing temp directory /tmp/reducer.hadoop.20230430.223102.103412...
○ [hadoop@ip-172-31-31-170 ~]$
○ [hadoop@ip-172-31-31-170 ~]$
```

# Executing the random generated file

## Input data

```
data.csv — hadoop [SSH: ec2-34-203-225-248.compute-1.amazonaws.com]
data.csv x  important commands.txt  part-00000
data.csv
1 0.48408779575496697, 0.25972515887359005, 0.7384829263040791, 0.7656350029041489, 0.09383904588327223, 0.16417557457954912, 0.4554982164674023, 0.8919094411770054
2 0.6570309057016853, 0.6061037310175563, 0.6612629898100094, 0.7931351111110198, 0.1787196667439932, 0.8648236474538414, 0.41038683188929803, 0.3528583119585466, 0
3 0.3231231841227039, 0.5094526918865543, 0.5207772986716622, 0.07715091920441564, 0.20489972353910113, 0.7192665506531439, 0.1268916016826802, 0.5743549718175202, 0
4 0.638004406411156, 0.8393886704194433, 0.3124420489657299, 0.9343266424870781, 0.722607800373723, 0.26023341864956306, 0.3905505917291626, 0.4350181219363083, 0
5 0.005517499557258887, 0.746964778793542, 0.30139408346199903, 0.5207350052565705, 0.3995309779101005, 0.2894951920474581, 0.9418900147537584, 0.9877894493701644, 0
6 0.37181547161418893, 0.41039985793457145, 0.23694257532408203, 0.6769355241624754, 0.3946822569904217, 0.53216130879877797, 0.6544267799370531, 0.6488099061810764, 0
7 0.4193938897247491, 0.5969597671653246, 0.3652785373187919, 0.10302877578996072, 0.7081400874676048, 0.14386563702655015, 0.46374418043428745, 0.27307685437122575
8 0.8378894295709511, 0.2014893000437047, 0.9491569698512646, 0.8408127092585804, 0.8129174989410972, 0.7538173335530762, 0.1655022644219446, 0.1249074889129611, 0.3
9 0.10053906240674748, 0.24499069284083974, 0.7535069218588585, 0.3496736357341721, 0.8159901438424935, 0.7785221009289696, 0.437789790631017, 0.9831115199290181, 0
10 0.7013510346930373, 0.8035333268496032, 0.4456406411587944, 0.40649484696143656, 0.9430083167613325, 0.9986818241900888, 0.7006111708718641, 0.4365187997426728, 0
11 0.1842964867264575, 0.33902895622881974, 0.7735650905799026, 0.26310884317110084, 0.8346935421607337, 0.7567832080214694, 0.532918194821051, 0.9834911647150223, 0
12 0.3140236141891055, 0.6834193075927283, 0.4014315014706583, 0.5774296185864305, 0.14047105034388774, 0.6936250604842009, 0.9044112391501469, 0.7161375571781503, 0
13 0.9840653042681918, 0.662374841893772, 0.38780703796739335, 0.4901489576260015, 0.9411954057066486, 0.6784392759394178, 0.39067566235161855, 0.8911117130033077, 0
14 0.8160434500605407, 0.3517395921187454, 0.8670732124378545, 0.22553013541858835, 0.6138703458983761, 0.14750726041547357, 0.6334414600629614, 0.9189897250692096, 0
15 0.16751008592985672, 0.829845079169215, 0.436359218142632, 0.635897841888343, 0.5885070945688087, 0.4473390146397489, 0.2207309441123524, 0.057397330896743814, 0.5
16 0.419198637438728, 0.27018110517518434, 0.42225178030196464, 0.7904567991215095, 0.17978133670658936, 0.40632745910221146, 0.41954967637093243, 0.4003666624800142
17 0.43584905137339014, 0.20868362461895418, 0.11037694168847734, 0.4577911275301655, 0.721347941271426, 0.020937278465378406, 0.910051236478965, 0.12086880516410614
18 0.18015123940080878, 0.8542591410166036, 0.6879997690682614, 0.21577177350662724, 0.5181913983489804, 0.060936582819478446, 0.44363226399979616, 0.2779135735636802
19 0.16924274082011714, 0.3186530035827295, 0.05261436732114322, 0.40376793743287265, 0.16954247579456594, 0.7932804148520181, 0.7493783624093738, 0.9387490100874216
20 0.25830294190794667, 0.3421800569126834, 0.8492260286440375, 0.8024321544424228, 0.5085630393945436, 0.396904788046643, 0.5761463883166561, 0.09414930482076866, 0
21 0.040464914512503114, 0.7162829807590103, 0.5242664553679217, 0.9194503197774319, 0.12365101332025696, 0.9815705788576594, 0.8813063423396718, 0.2935216550790887, 0
22 0.25851324822457733, 0.02382740717488685, 0.3818158411413186, 0.7757250985718254, 0.40642340030477886, 0.3297189026350338, 0.10918154784099343, 0.6978865805312993
23 0.9887278504155286, 0.7218844107538216, 0.7209269397586443, 0.6630029152960866, 0.3325558303702575, 0.6754329285914371, 0.8671716446748303, 0.9711427047111652, 0.3
24 0.304607823911914, 0.4695427046758731, 0.4717673334586796, 0.2806001725454108, 0.9927526974559565, 0.30167478895179034, 0.7384927095038089, 0.7011870165209229, 0.8
25 0.32353789723834603, 0.703880610484651, 0.15794340743402657, 0.188924321550619, 0.7800970336283598, 0.9908315502566141, 0.5983164097254577, 0.07127899842433827, 0
26 0.897118266769053, 0.3082076628643494, 0.05122138960903999, 0.5966173421270544, 0.5163228605965935, 0.4723911030232716, 0.6942695499720344, 0.5638284608950418, 0
27 0.2701753502864511, 0.5460579784706182, 0.7266988436193494, 0.9830086234393067, 0.5471534109502088, 0.14467036427397206, 0.7146646096904088, 0.6157049789842072, 0
28 0.8650114208117473, 0.17158564630584383, 0.29121186819379974, 0.8340758631205522, 0.6492934515354304, 0.1518781794468118, 0.8414192647987693, 0.5714257747875076, 0
29 0.9321866685226696, 0.4742185027816968, 0.631395383646845, 0.905968743778733, 0.4483877681026086, 0.13343454474633198, 0.5544923947219799, 0.21732592350055546, 0.8
30 0.61017707474732055, 0.4740315692317554, 0.20452305993591302, 0.9184201179305733, 0.7096371672001134, 0.7519102566933262, 0.91287095614597, 0.08396510153487513, 0.2
31 0.1147580692901431, 0.91732163461008597, 0.8535118917478923, 0.7213115149368384, 0.757040213910103, 0.6524207367396871, 0.6670221087030064, 0.9648438597617551, 0.96
32 0.3109974199399359, 0.1792053661870815, 0.3520835001431989, 0.8235450501277296, 0.9051805377126637, 0.2508091983075149, 0.020853846089596617, 0.22605897854344537, 0
33 0.07203152698883442, 0.9207145379771893, 0.3695012626683829, 0.025446565871965365, 0.7223944612494099, 0.8977511395119867, 0.49129732202037006, 0.8833456336268576
34 0.21490497887540094, 0.5992900698057329, 0.9887407643814257, 0.23038745150295173, 0.3811636251379257, 0.017934222832766133, 0.6158910288965683, 0.4157016877519681
35 0.41484436713615214, 0.9688240880711783, 0.8048548177255719, 0.8573279955995593, 0.09845082482136736, 0.4791965115468275, 0.5153478586086297, 0.5596793773657796, 0
36 0.5551357113280001, 0.598286644761441, 0.7176955650633126, 0.22551372013428483, 0.5697583085015903, 0.23538036832782094, 0.1564441491126326, 0.474689243084932, 0.2
37 0.7733001578840414, 0.3311853685696038, 0.8853094852005816, 0.3152237578137096, 0.4420150375964851, 0.10066305589113278, 0.6455794757371752, 0.7310684475228202, 0
38 0.3922610193314723, 0.7733625345407111, 0.9668299622151058, 0.17272557505771846, 0.43151375280473026, 0.8403236517798381, 0.3975324808336478, 0.10934247938091768, 0
39 0.1271075475607295, 0.10656003047367912, 0.647528692366261, 0.2280089584862192, 0.7512529158689479, 0.45752971390259234, 0.32324479048669896, 0.7502875699267101, 0
40 0.9878410979980707, 0.04787627790932303, 0.48753975238099223, 0.6134458280559026, 0.39286873103432685, 0.9762360659322457, 0.6329207064595553, 0.542623590127919, 0
41 0.5338945680071531, 0.6645662216007721, 0.252588383205375, 0.4555292360643145, 0.5537489823746595, 0.29577575943716705, 0.47976667847740273, 0.9800748500229544, 0
42 0.24509104663350723, 0.2169739447521527, 0.7166347351070528, 0.6082956661755968, 0.44238065073273954, 0.6239185218305349, 0.30061569400400545, 0.8442378604474173, 0
43 0.6645575050086018, 0.79610740207034566, 0.4486430030151041, 0.78474787108540524, 0.7025551555620628, 0.30712186457420916, 0.80788075721103557, 0.6322276367647032, 0
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
bash + - x
```



# Output data

data.csv

important commands.txt

part-00000

mapper.py 1

part-00000

```
1  "0.6044676336944209  0.1894546555809693  0.7163835760649801  0.6086977102383268  0.0688621935652105  0.24566813108249963  0.9281392585046088  0.69704054054
2  "0.6311487147652488  0.40853311918637014  0.7724341649870091  0.6666399275960374  0.3011002784857263  0.03053327670968764  0.6753126960081373  0.3390039888
3  "0.8658776307118707  0.9134754116008316  0.4510776240447043  0.23305605475702096  0.18874008379726592  0.13363670512666848  0.7051061601064439  0.23589370
4  "0.5098915030559537  0.4391792227008712  0.7809252999911804  0.3077694402703254  0.5228176028746254  0.16989237805091362  0.16046347546752748  0.3031060577
5  "0.3250623258920632  0.7409716293162474  0.08063440979094649  0.7475623913810239  0.05751621620916125  0.5456587482367001  0.7773587965873408  0.7214853505
6  "0.7728038308208586  0.864030650634218  0.10156094954484951  0.28489332786513333  0.30191656104550424  0.5771575306293899  0.18946319931285638  0.663965924
7  "0.8459536674113179  0.5561939128679022  0.8932711505906193  0.7794945348419589  0.7790183444627046  0.07942250220814917  0.07006030818145226  0.6360256201
8  "0.650730747273292  0.9447880687188789  0.3142323741683072  0.13010071026135794  0.49932109210336106  0.20643395006421117  0.49517810494455705  0.56242324
9  "0.2027156088822809  0.8763429958702177  0.11613674017231135  0.08431069013056036  0.9228086593982763  0.22569864071240764  0.8583062156542034  0.10722496
10 "0.7329979222202644  0.010307006647438444  0.13045935357751393  0.04168554783488043  0.9045010007942368  0.28943045377132703  0.9922092038646618  0.896556
11 "0.5472787421490416  0.4911290934543048  0.7378135067687859  0.6070783319393428  0.30505266615748405  0.9038246082636817  0.6858947967159352  0.9609152353
12 "0.8830012649898382  0.3554421383938442  0.836387803231762  0.522145645689498  0.5018427856754203  0.8929583751996466  0.4526797884949395  0.8774102125731
13 "0.09004189146370023  0.6399755068348327  0.8720230295328166  0.2506044880631333  0.4211037425152022  0.27307904106612946  0.7726664633897941  0.277536744
14 "0.9827427408306036  0.5038175916505703  0.021619714029426973  0.7366497698551115  0.011344891627495968  0.29540365525522816  0.5581588597107174  0.319292
15 "0.09232369907086024  0.9378839686057018  0.16750906670775711  0.464949149259152  0.7676767738289589  0.13074054189308804  0.7446606029895865  0.31824208
16 "0.9671786328435764  0.6387431488328457  0.38158505738232376  0.5506494589369364  0.09551418246930243  0.5540895133583787  0.10252426678193971  0.17200976
17 "0.49552236129103944  0.4201479147046693  0.33531555034007166  0.766650790824644  0.8754365206848883  0.6016641709108601  0.43883095008049515  0.882607289
18 "0.572282536668324  0.49077805017866605  0.7014159742109859  0.7665279720760321  0.6334740046190804  0.4758159800752278  0.6871401217546806  0.80216594024
19 "0.619288478925191  0.5253349281305664  0.07806583833787084  0.30136707834239296  0.4217786407281269  0.5472723663504984  0.04434576635727716  0.3475136
20 "0.44267441678956576  0.841528957864267  0.6249574335286993  0.9404608309441524  0.33645928145788806  0.9139011391711716  0.3844230942865121  0.2381646176
21 "0.6606108968587764  0.7751302228586632  0.2863325328543782  0.1337355073129448  0.41578867608994996  0.9412986521965464  0.7339849341020442  0.4023066384
22 "0.9539062314163252  0.07343849339114217  0.6852492409062442  0.26666345639411515  0.43252841700475086  0.20861016992001957  0.7179436894032762  0.9797846
23 "0.5150038863736662  0.4949634607184299  0.6798781674948636  0.7795404556338791  0.31365330021195237  0.6401213416495634  0.8321047924723992  0.98090054811
24 "0.986667711054282  0.14213975784729382  0.86012273274668  0.188308677192178  0.3853297596109083  0.5343752035974859  0.92674162693992885  0.437535450924
25 "0.07447035572221228  0.8695170733292734  0.6971760194657914  0.7508399929351891  0.35398807385848974  0.41536118664952815  0.28640852968248287  0.8251147
26 "0.3020857513048299  0.13763968020953854  0.8130826344050944  0.763467099208927  0.14138980808072656  0.34356470638321464  0.9111210600643545  0.275698090
27 "0.411789808413361  0.13221082319537614  0.748125436736499  0.7278945808611112  0.773037231749792  0.4442608697758148  0.803255662245502  0.2232004824080
28 "0.6753216237532312  0.15034282864399406  0.24408232779934225  0.3140388520296842  0.6267049135054076  0.932495070621422  0.07846121592847888  0.481188723
29 "0.05662182141150918  0.7585834879080003  0.7859905867207828  0.7260093229093327  0.011981112506471825  0.26001420245841866  0.5949325147712878  0.9550624
30 "0.18577539741017868  0.562927164676851  0.9661539937045726  0.4026155205854136  0.6505618056231474  0.05539713121220702  0.7699125716752886  0.3020994905
31 "0.3319844703372471  0.6760541414755731  0.368155643542209  0.9880484492767356  0.7175092994838004  0.04743847814404978  0.7595416997216591  0.745512751091
32 "0.5496416327256115  0.24736963597235628  0.9721621026778354  0.4838583290586539  0.1866909663218851  0.7454309528903694  0.8096335437877022  0.39064071211
33 "0.3596563302947209  0.8047451107080295  0.09309070612062387  0.8365099598397921  0.32032937984626597  0.3393014364774247  0.7912729536398142  0.575683874
34 "0.5454064518680408  0.9839217832313065  0.0928247670171834  0.8600379756703789  0.2728567329791889  0.1782009801151364  0.7718434318927582  0.08685779882
35 "0.8012688515194561  0.25618409115667187  0.003745265805951248  0.21313745160021713  0.8527367958738742  0.7095521843013955  0.9081568589239934  0.6023216
36 "0.5419693317733045  0.031174600803983266  0.5817035733068968  0.7495725676053802  0.03736397616734133  0.25059283305460545  0.37137052132625514  0.756220
37 "0.3241340776772551  0.8001658288217905  0.8257870911156361  0.6154574898698928  0.7630085820289141  0.060805919631545136  0.30808678468542794  0.87467878
```

PROBLEMS 1

OUTPUT

DEBUG CONSOLE

TERMINAL

PORTS

Reduce input records=1000

Reduce output records=1000

Reduce shuffle bytes=157563

Shuffled Maps =4

Spilled Records=2000

Total committed heap usage (bytes)=1595932672

Virtual memory (bytes) snapshot=17875226624

Shuffle Errors

BAD\_ID=0

CONNECTION=0

IO\_ERROR=0

WRONG\_LENGTH=0

WRONG\_MAP=0

WRONG\_REDUCE=0

job output is in hdfs:///final\_output2

Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/reducer.hadoop.20230430.233528.170978...

Removing temp directory /tmp/reducer.hadoop.20230430.233528.170978...

o [hadoop@ip-172-31-25-78 ~]\$

o [hadoop@ip-172-31-25-78 ~]\$

Ln 1, Col 1

Spaces: 4

UTF-8

LF

Plain Text

## Job run screenshots.

Execution time: 38s

```
PROBLEMS 1 OUTPUT DEBUG CONSOLE TERMINAL PORTS
[hadoop@ip-172-31-25-78 ~]$ time python mapper.py -r hadoop hdfs:///user/hadoop/data.csv --output-dir hdfs:///output2 --jobconf mapreduce.job.reduces=1 && python reducer.py -r hadoop
hdfs:///output2/* --output-dir hdfs:///final_output2
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/mapper.hadoop.20230430.233422.906167
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/mapper.hadoop.20230430.233422.906167/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/mapper.hadoop.20230430.233422.906167/files/
Running step 1 of 1...
packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-4.jar] /tmp/streamjob4203084412314391749.jar tmpDir=null
Connecting to ResourceManager at ip-172-31-25-78.ec2.internal/172.31.25.78:8032
Connecting to Application History server at ip-172-31-25-78.ec2.internal/172.31.25.78:10200
Connecting to ResourceManager at ip-172-31-25-78.ec2.internal/172.31.25.78:8032
Connecting to Application History server at ip-172-31-25-78.ec2.internal/172.31.25.78:10200
Loaded native op 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_1682896939926_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_1682896939926_0004
The url to track the job: http://ip-172-31-25-78.ec2.internal:20888/proxy/application_1682896939926_0004/
Running job: job_1682896939926_0004
Job job_1682896939926_0004 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_1682896939926_0004 completed successfully
Output directory: hdfs:///output2
Counters: 49
  File Input Format Counters
    Bytes Read=244638
  File Output Format Counters
    Bytes Written=208738
  File System Counters
    FILE: Number of bytes read=156248
    FILE: Number of bytes written=1434587
    FILE: Number of large read operations=0
    FILE: Number of read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=245090
    HDFS: Number of bytes written=208738
    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
    Launched map tasks=4
    Launched reduce tasks=1
    Total megabyte-milliseconds taken by all map tasks=58684416
    Total megabyte-milliseconds taken by all reduce tasks=12438528
```

```
Job job_1682896939926_0004 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_1682896939926_0004 completed successfully
Output directory: hdfs:///output2
Counters: 49
  File Input Format Counters
    Bytes Read=244638
  File Output Format Counters
    Bytes Written=208738
  File System Counters
    FILE: Number of bytes read=156248
    FILE: Number of bytes written=1434587
    FILE: Number of large read operations=0
    FILE: Number of read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=245090
    HDFS: Number of bytes written=208738
    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
    Launched map tasks=4
    Launched reduce tasks=1
    Total megabyte-milliseconds taken by all map tasks=58684416
    Total megabyte-milliseconds taken by all reduce tasks=12438528
    Total time spent by all map tasks (ms)=38206
    Total time spent by all maps in occupied slots (ms)=1833888
    Total time spent by all reduce tasks (ms)=4049
    Total time spent by all reduces in occupied slots (ms)=388704
    Total vcore-milliseconds taken by all map tasks=38206
    Total vcore-milliseconds taken by all reduce tasks=4049
  Map-Reduce Framework
    CPU time spent (ms)=4330
    Combine input records=0
```

```

Total HDFS milliseconds taken by all Reduce tasks=4843
Map-Reduce Framework
  CPU time spent (ms)=4330
  Combine input records=0
  Combine output records=0
  Failed Shuffles=0
  GC time elapsed (ms)=916
  Input split bytes=452
  Map input records=1000
  Map output bytes=209738
  Map output materialized bytes=156759
  Map output records=1000
  Merged Map outputs=4
  Physical memory (bytes) snapshot=2073980928
  Reduce input groups=1
  Reduce input records=1000
  Reduce output records=1000
  Reduce shuffle bytes=156759
  Shuffled Maps =4
  Spilled Records=2000
  Total committed heap usage (bytes)=1670905856
  Virtual memory (bytes) snapshot=17882116096
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
job output is in hdfs:///output2
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/mapper.hadoop.20230430.233422.906167...
Removing temp directory /tmp/mapper.hadoop.20230430.233422.906167...

real    1m5.387s
user    0m38.330s
sys     0m2.611s

```

```
PROBLEMS 1 OUTPUT DEBUG CONSOLE TERMINAL PORTS
Counters: 49
File Input Format Counters
  Bytes Read=247890
File Output Format Counters
  Bytes Written=218738
File System Counters
  FILE: Number of bytes read=157555
  FILE: Number of bytes written=1440663
  FILE: Number of large read operations=0
  FILE: Number of read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=248334
  HDFS: Number of bytes written=218738
  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
  Launched map tasks=4
  Launched reduce tasks=1
  Total megabyte-milliseconds taken by all map tasks=61214208
  Total megabyte-milliseconds taken by all reduce tasks=12923904
  Total time spent by all map tasks (ms)=39853
  Total time spent by all maps in occupied slots (ms)=1912944
  Total time spent by all reduce tasks (ms)=4207
  Total time spent by all reduces in occupied slots (ms)=403872
  Total vcore-milliseconds taken by all map tasks=39853
  Total vcore-milliseconds taken by all reduce tasks=4207
Map-Reduce Framework
  CPU time spent (ms)=4820
  Combine input records=0
  Combine output records=0
  Failed Shuffles=0
  GC time elapsed (ms)=879
  Input split bytes=444
  Map input records=1000
  Map output bytes=218738
  Map output materialized bytes=157563
  Map output records=1000
  Merged Map outputs=4
  Physical memory (bytes) snapshot=2043240448
  Reduce input groups=1
  Reduce input records=1000
  Reduce output records=1000
  Reduce shuffle bytes=157563
  Shuffled Maps =4
  Spilled Records=2000
  Total committed heap usage (bytes)=1595932672
  Virtual memory (bytes) snapshot=17875226624
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
job output is in hdfs:///final_output2
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mr/job/reducer.hadoop.20230430.233528.170978...
Removing temp directory /tmp/reducer.hadoop.20230430.233528.170978...
[hadoop@ip-172-31-75-78 ~]$
```

## Executing the Spotify dataset from Kaggle

The EMR cluster starts reconnecting for the bigger size files. Soon after the cluster encounters it, it starts disconnecting and exits from backend because the file that was just being uploaded filled up the complete memory on the instance, though the cluster is still in the waiting status in the AWS console.

Therefore, I used random file generator option to create a file within the EC2 to opt for smaller size which could be stored on the EMR cluster without causing issues(Like the spotify dataset which was about 6GB which filled up the EBS volume and caused system reboot)

Please find the below screenshots for your reference.

```
important commands.txt
1 pip install mrjob
2
3
4
5 hdfs dfs -put spotify.csv /user/hadoop
6
7 time python mapper.py -r hadoop hdfs:///user/hadoop/data.csv --output-dir hdfs:///mainoutput2 --jobconf mapreduce.job.reduces=1 && python reducer.py -r hadoop hdfs:///mainoutput2 --jobconf mapreduce.job.reduces=1
```

```
(from line 55 of hdfs:///tmp/hadoop-yarn/staging/history/done_intermediate/hadoop/job_1682892957887_0004-1682895418486-hadoop-streamjob7437597255060536547.jar-1682895477780-3-0-FAILED-default-1682895425701.jhist)

Step 1 of 1 failed: Command '['/usr/bin/hadoop', 'jar', '/usr/lib/hadoop-mapreduce/hadoop-streaming.jar', '-files', 'hdfs:///user/hadoop/tmp/mrjob/mapper.hadoop.20230430.225629.880476/files/wd/mapper.py#mapper.py,hdfs:///user/hadoop/tmp/mrjob/mapper.hadoop.20230430.225629.880476/files/wd/mrjob.zip#mrjob.zip,hdfs:///user/hadoop/tmp/mrjob/mapper.hadoop.20230430.225629.880476/files/wd/setup-wrapper.sh#setup-wrapper.sh', '-D', 'mapreduce.job.reduces=0', '-D', 'mapreduce.job.reduces=1', '-input', 'hdfs:///user/hadoop/spotify.csv', '-output', 'hdfs:///mainoutput', '-mapper', '/bin/sh -ex setup-wrapper.sh python3 mapper.py --step-num=0 --mapper']' returned non-zero exit status 256.

real    2m1.810s
user    1m4.101s
sys      0m4.422s
[hadoop@ip-172-31-31-170 ~]$ python --version
Python 3.7.16
[hadoop@ip-172-31-31-170 ~]$ python3 filegen.py
```

```
output-dir hdfs:///mainoutput2 --jobconf mapreduce.job.reduces=1 && python reducer.py -r hadoop hdfs:///mainoutput2/* --output-dir hdfs:///final_mainoutput2
```

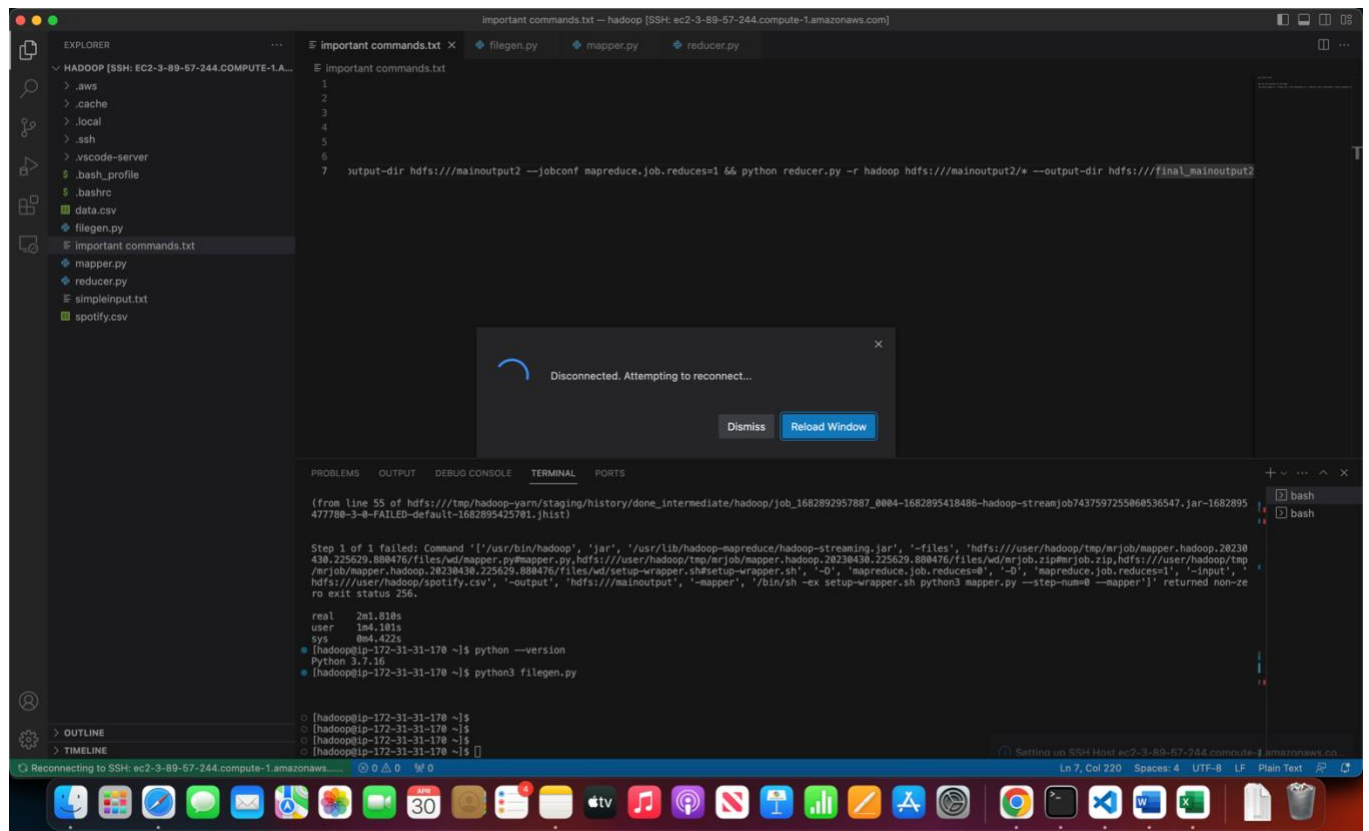
```
(from line 55 of hdfs:///tmp/hadoop-yarn/staging/history/done_intermediate/hadoop/job_1682892957887_0004-1682895418486-hadoop-streamjob7437597255060536547.jar-1682895477780-3-0-FAILED-default-1682895425701.jhist)

Step 1 of 1 failed: Command '['/usr/bin/hadoop', 'jar', '/usr/lib/hadoop-mapreduce/hadoop-streaming.jar', '-files', 'hdfs:///user/hadoop/tmp/mrjob/mapper.hadoop.20230430.225629.880476/files/wd/mapper.py#mapper.py,hdfs:///user/hadoop/tmp/mrjob/mapper.hadoop.20230430.225629.880476/files/wd/mrjob.zip#mrjob.zip,hdfs:///user/hadoop/tmp/mrjob/mapper.hadoop.20230430.225629.880476/files/wd/setup-wrapper.sh#setup-wrapper.sh', '-D', 'mapreduce.job.reduces=0', '-D', 'mapreduce.job.reduces=1', '-input', 'hdfs:///user/hadoop/spotify.csv', '-output', 'hdfs:///mainoutput', '-mapper', '/bin/sh -ex setup-wrapper.sh python3 mapper.py --step-num=0 --mapper']' returned non-zero exit status 256.

real    2m1.810s
user    1m4.101s
sys      0m4.422s
[hadoop@ip-172-31-31-170 ~]$ python --version
Python 3.7.16
[hadoop@ip-172-31-31-170 ~]$ python3 filegen.py
```

Disconnected. Attempting to reconnect...

Dismiss Reload Window



## Test cases:

### Case 1: Empty file check.

```
[hadoop@ip-172-31-17-213 ~]$ time python mapper.py -r hadoop hdfs:///user/hadoop/data.csv --output-dir hdfs:///output5 --jobconf mapreduce.job.reduces=1 && python reducer.py -r hadoop hdfs:///output5/* --output-dir hdfs:///final_output5
File is empty!

real    0m0.212s
user    0m0.174s
sys     0m0.024s
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
```

### Case 2: File size check to avoid EMR reboot.

```
[hadoop@ip-172-31-17-213 ~]$ time python mapper.py -r hadoop hdfs:///user/hadoop/empty.csv --output-dir hdfs:///output --jobconf mapreduce.job.reduces=1 && python reducer.py -r hadoop hdfs:///output/* --output-dir hdfs:///final_output
The File size acceptable, Pushed to HDFS.
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...
```



### Case 3: There no commas within the given file.

My test case fails here. The additional code that I tried to add does not seem to get the count. However, normally if there are no commas then there won't be any replace, and new file generated content will be same as old file. In the absence of additional code for this part, the job runs successfully without making any modifications. The below screenshots reflects the run with additional code.

```
mapper.py  sizetest.py  to be addedmapper.py  important commands.txt  impmapper.py  reducer.py  NoCommaFile.txt  part-00000
@ part-00000
1 Error: This file has no commas to replace with space
2 Error: This file has no commas to replace with space
3 Error: This file has no commas to replace with space
4 Error: This file has no commas to replace with space
5 null
"1122331341343451255113242555231141325144342214214513432413455342334125524542545351432231351251541133311513511422424221544315154124545312522312321411424511
312231545324325114123513112553411333411234151214142152232534341155312452355441135215255551114241253133524223233155321312445142441444235224523323535232442
334342335454523131341131245323253141513551241135131433413434132453254122445234453133244251325554123315331341345134153424123342511142414432123423
3212531231335123233155215551331423422223345315113442231444255332254434135122555542112343341244435311554334251155113334444131452421121341222523543525345253
414532212543342141521353511214354115534542214134522552351522313543144522532231425111421445233521542431422513445532243123255512311235424212113142441121245
1122514133111542511154543433155141413213443414533425134554521143143113545235154423552454311225251314541211245444312444425453231233321232552155544514421513
5251531521112454235121523521414445333354313552524334251212544123225225124532215455235142131533453215211211543443312555334213253232552245252311155142113141
542122512542522213124421313443355142154241124233352412433255513151123423422234511552431511452451112543331213354144144544515123523523123424343522454445311
323252532135512253554414232125154544124451511325553554225124122415533531512434242234233442341555511343413551524443241121123144155245553431525512223414
51424145115345425353313322245415225412251151521531425355355155452333122344122525342122452411252355413423443211242312244112334413534232445311542343422111
24235414424115254532411214242133121125555325324241143343244442451335435354421353423553253512445255254555521431453351444323355525452241244132231255535154
21414445335533351214421124313541525243412512424431135344425521225453425545444514245342322433212222553344353243253312515232322525422442421314345255541311
155145145155332255233421545245541532531213515521245415411212325223544143453554351442552524433455115454135421531251423144311244342314221443544453444543415
14222113353135114213424435532135314224532122542154334424321423232251115325534522135214112251453135125415312521211431343241431344544533223511543234551213512
4541221455544152314145212242232142135151552522154442253554533443234154332424345123155445144145341344314324453514253331134113451452215125254454152111544132
224515555535544252351332113242225145253442135444513444355241255255335242412442541355145454533511135232231123115154333412541434452543413521314144213552342
15421234331153321443511153152423224342114514432351452151231553551454452244432542553334543513515132511522315541142415322121554251142213353543515535153554553
115455444153344345451131415315144343155135532531555544154223514512355321221553424554351232424453345551312341232252515125441443355444231444243413231232132113
551335445415534152445411121245253545342132423421514233225543342134453131432233553244545225411222231455214143154552245123443441531254342214114542355344445152
2241213555153242245332145134132114453415422341211432412512345312551522513244425331141245533432234432135514142235342353522142522453341434434324531214342222
124413123213544515342353121324424232511542224125244443231231221142544511212453331141531532545413115425413422141214354215532251131215125215551531342213543
5521235124211534312451535144354124121144555442434342422123154411452321211212254424121224254252542241524435222515311312133324153113524443321454343224314353
2534421345342415153543131312244153311115545453413311524242215435552254414434322141151551423353554114333245143315441323114343134153512335551113534114334
4123531545143311152154421225543334332451344334345222345253431413253555132121353512312555421344451243414152345352245245152524355122315524152533252424253313
42343255323214535441244554523241515224112232135243514545412133243114351114222312423255523441154122223422545341452443433522412523211515543343442323111311312
35333413111154131343421153522324431543224532441241514144522252253342352351552344425545355112543222523442514113241522425423544212523123314411354533133344
5143535421525144113254154554414431333544132322131315125434335252131515134223114214154422312551314433335413244431132143432542523325153524345521333233215554
33245425444134143311331251124142134232542521453225352121235115531434454342255124324331452431135135332114524514542114213151255312135145454234415533513334
12523225151351544451423113341152221213433153133235121325123213222123354544415325452312113111133313125121544132215521245244243525534411454132323515435433544
123514444244453422314513534352414154152313514335111245515111454312245452551333514155141121351524343555343211341341354325253152522215533523135223533542413
3141542351334342315552525345441432454142335454124424554332225154523235253345441531123531552453552545442152244434144552254522445154542433453333151512535534
3355533123353554424343511115422121421551423212144253514135344315134524112542151531134143331422225231231222135233434223531423312552221111141245421532335
34242523331331445344323333535444113533115521121524543551411412241253413413255114145343544122134522134232441344345144111331151252455344555312152253121112543
5143514344442112453521441353333154132451521211432325144334412434512145224525542355451421442143222513221232552425331525123352432252434
41245341231124325143135311525242421221113352225144151223211531443142423221145322134541441423223234445522122211313345123413312532445335511223214522531244323
1355234542332445153211534542235514445345153321513333112342351434112343123412553542311521453443552545443131234322112142423452222115554515315114431251144242
425341151555133325224325343321431155254411321142153341541534534352134534111452314242342125351415415514225313441343445332453411243553552145235252242431
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
[hadoop@ip-172-31-17-213 ~]$
[hadoop@ip-172-31-17-213 ~]$ hdfs dfs -get /output60/part-00000
[hadoop@ip-172-31-17-213 ~]$
```



## **Conclusion:**

I was able to successfully convert the file format from Comma Separated Values (CSV) to Tab Separated Values (TSV). I was able to rectify majority of the issues that was reflecting in the draft file such as /t was reflecting in the output file rather than actual tab. Currently I am successfully able to export the file where the contents are separated by tabs (instead of /t).

## **Reference:**

- <https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html>
- [https://www.tutorialspoint.com/hadoop/hadoop\\_mapreduce.htm](https://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm)
- [https://medium.com/@dhammikasamankumara/what-is-hadoop-distributed-file-system-hdfs-36a3503f9c60#:~:text=Hadoop%20distributed%20file%20System%20\(HDFS\)%20splits%20the%20large%20data%20files,result%20in%20data%20being%20unavailable](https://medium.com/@dhammikasamankumara/what-is-hadoop-distributed-file-system-hdfs-36a3503f9c60#:~:text=Hadoop%20distributed%20file%20System%20(HDFS)%20splits%20the%20large%20data%20files,result%20in%20data%20being%20unavailable)