cat big.txt | ./mapper.py

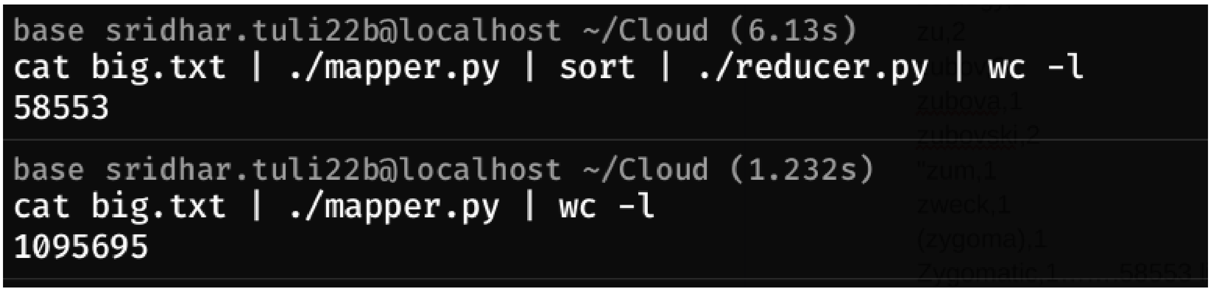
OUTPUT:

```
print,1
i,,1
c,,1
ord(c),,1
big[max(0,,1
i-10):min(n,,1
i+10)],1
s.add(c),1
print,1
s,1
print,1
[ord(c),1
for,1
c,1
In,1……. 1095695 lines
```

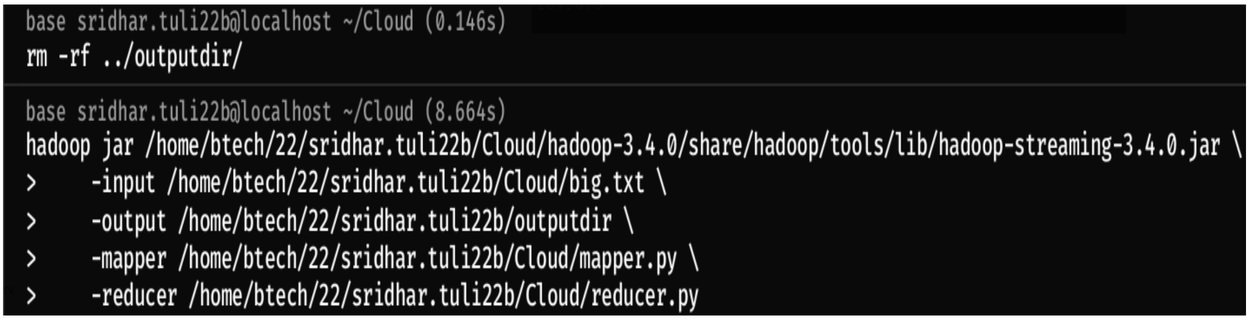
cat big.txt | ./mapper.py | sort | ./reducer.py

OUTPUT:

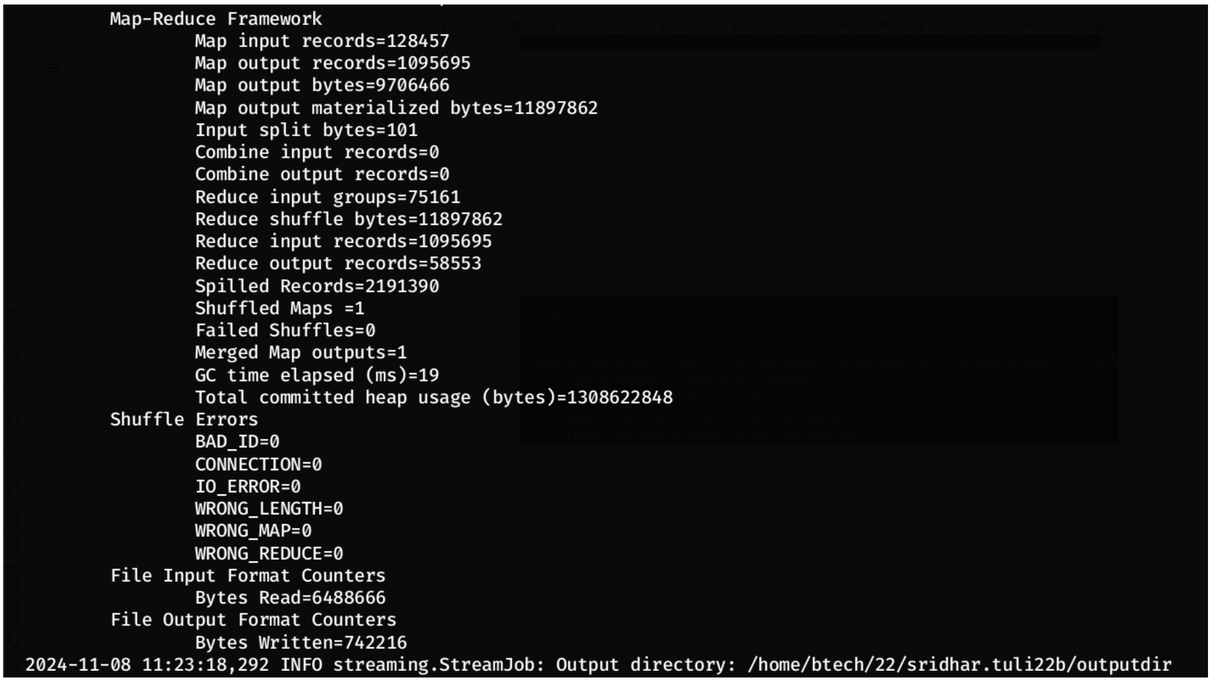
```
znaim.",1
znamenka.,1
zone,23
zone.,2
zone--not,1
zoology,1
zu,2
zubov,2
zubova,1
zubovski,2
"zum,1
zweck,1
(zygoma),1
Zygomatic,1…….58553 lines
```



```
base sridhar.tuli22b@localhost ~/Cloud (6.13s)
cat big.txt | ./mapper.py | sort | ./reducer.py | wc -l
58553

base sridhar.tuli22b@localhost ~/Cloud (1.232s)
cat big.txt | ./mapper.py | wc -l
1095695
```

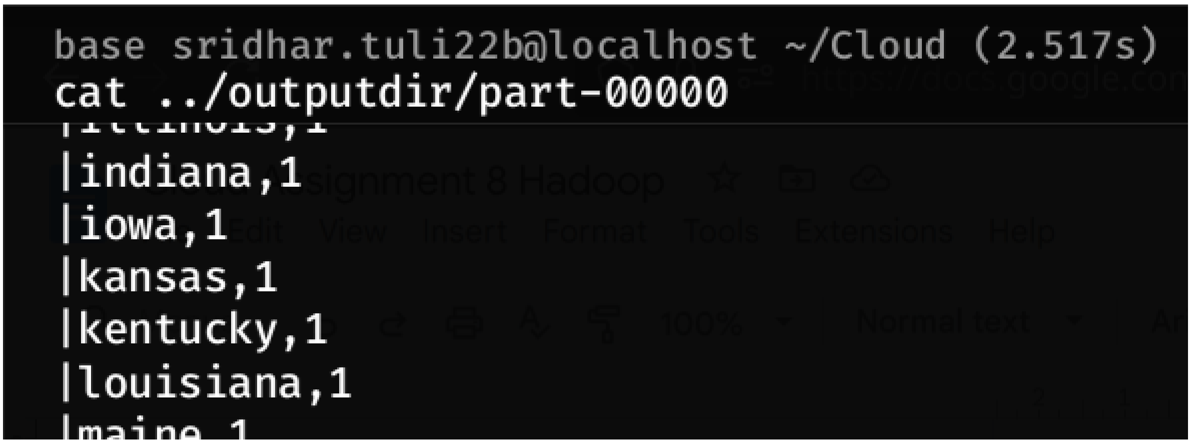


```
base sridhar.tuli22b@localhost ~/Cloud (0.146s)
rm -rf ../outputdir/

base sridhar.tuli22b@localhost ~/Cloud (8.664s)
hadoop jar /home/btech/22/sridhar.tuli22b/Cloud/hadoop-3.4.0/share/hadoop/tools/lib/hadoop-streaming-3.4.0.jar \
>     -input /home/btech/22/sridhar.tuli22b/Cloud/big.txt \
>     -output /home/btech/22/sridhar.tuli22b/outputdir \
>     -mapper /home/btech/22/sridhar.tuli22b/Cloud/mapper.py \
>     -reducer /home/btech/22/sridhar.tuli22b/Cloud/reducer.py
```



```
        Map-Reduce Framework
                Map input records=128457
                Map output records=1095695
                Map output bytes=9706466
                Map output materialized bytes=11897862
                Input split bytes=101
                Combine input records=0
                Combine output records=0
                Reduce input groups=75161
                Reduce shuffle bytes=11897862
                Reduce input records=1095695
                Reduce output records=58553
                Spilled Records=2191390
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=19
                Total committed heap usage (bytes)=1308622848
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=6488666
        File Output Format Counters
                Bytes Written=742216
2024-11-08 11:23:18,292 INFO streaming.StreamJob: Output directory: /home/btech/22/sridhar.tuli22b/outputdir
```

OUTPUT

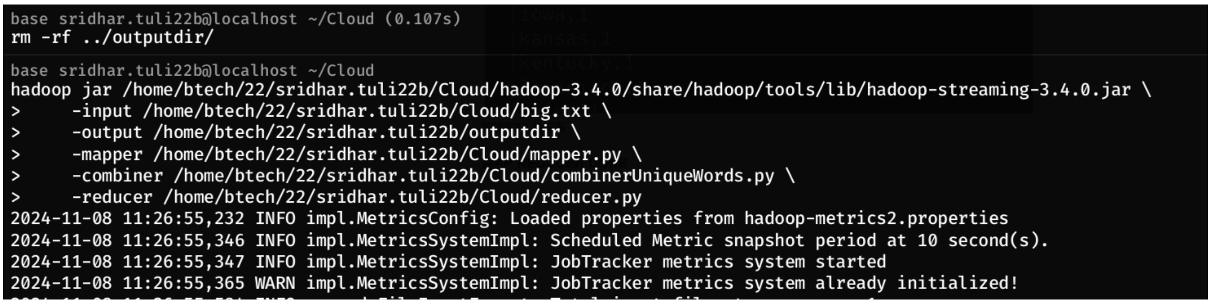


```
base sridhar.tuli22b@localhost ~/Cloud (2.517s)
cat ../outputdir/part-00000
|illinois,1
|indiana,1
|iowa,1
|kansas,1
|kentucky,1
|louisiana,1
|maine,1
```

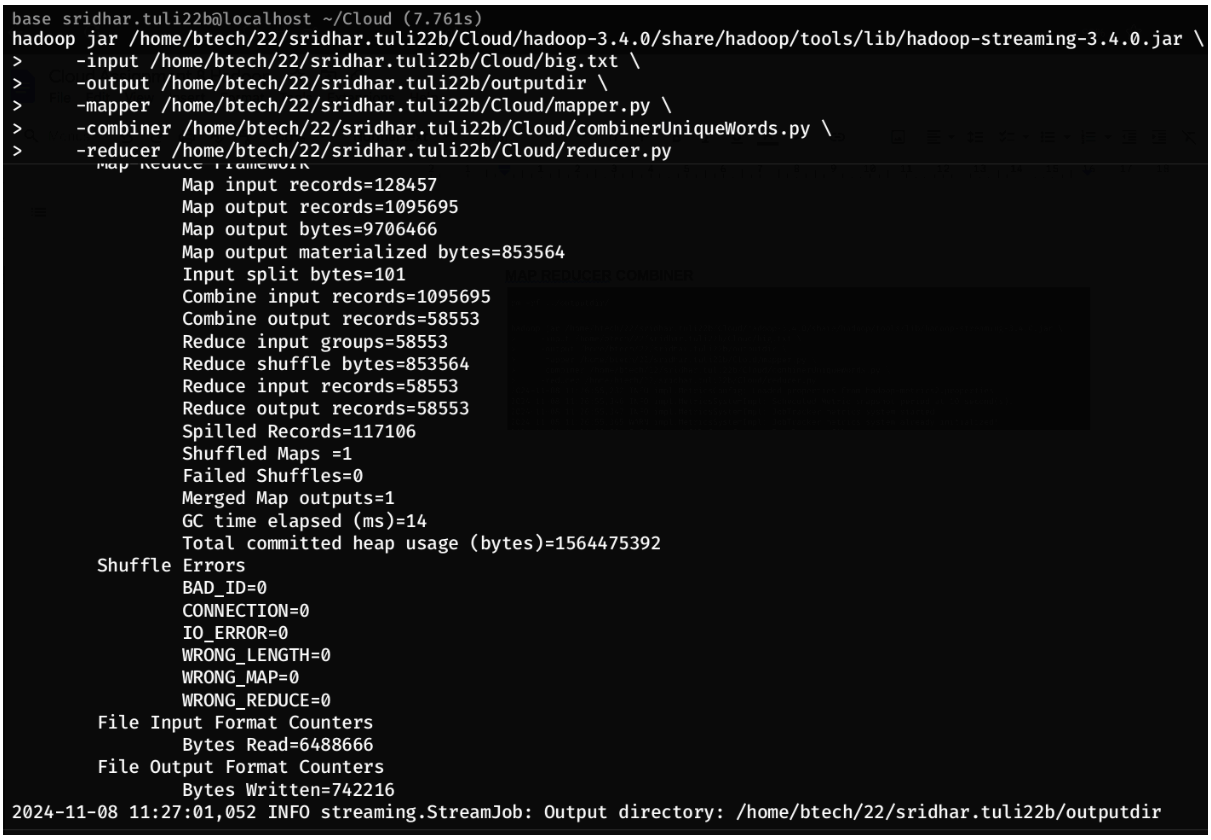
58553 LINES

MAP REDUCER DONE

MAP REDUCER COMBINER

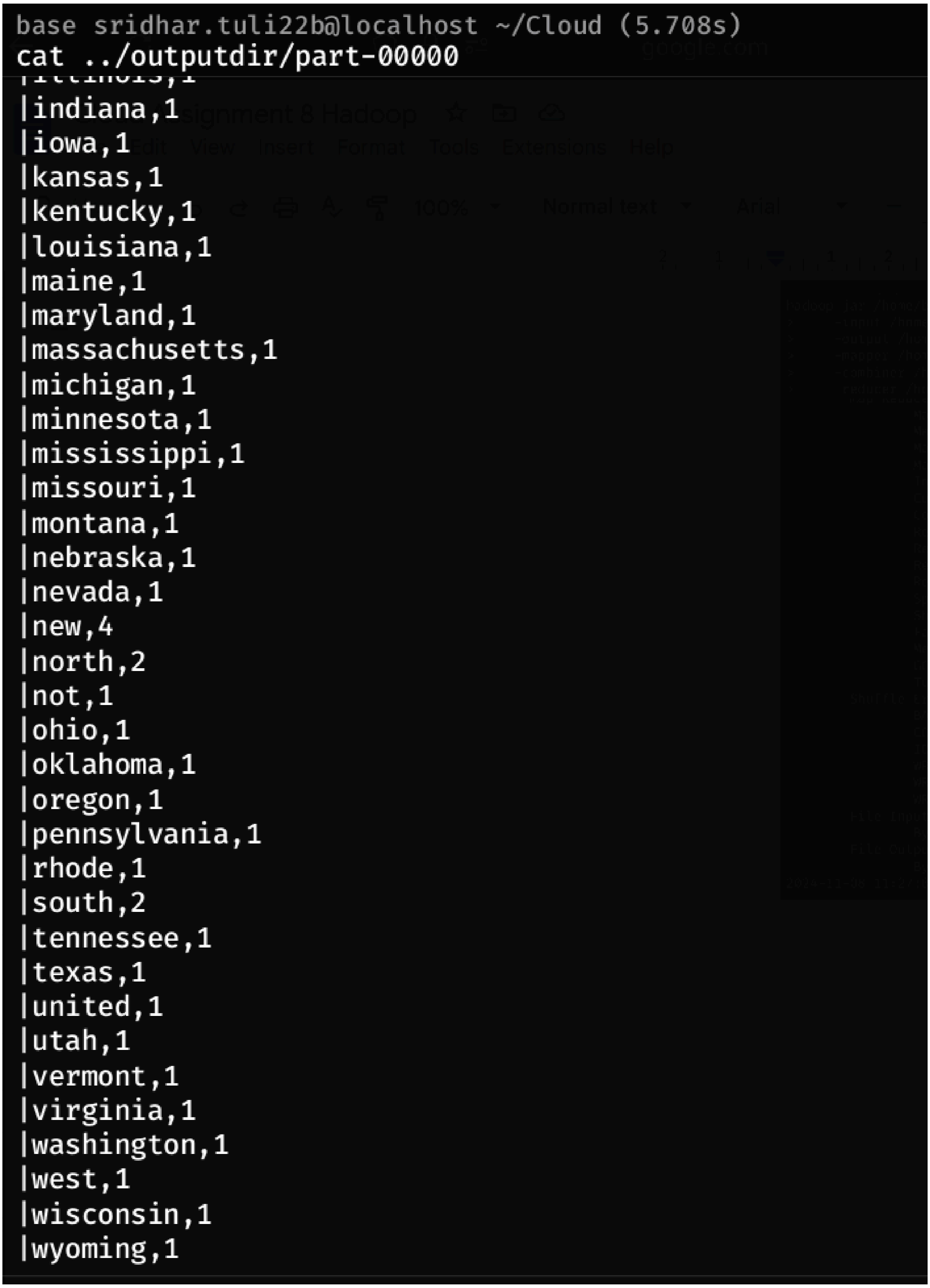


```
base sridhar.tuli22b@localhost ~/Cloud (0.107s)
rm -rf ../outputdir/

base sridhar.tuli22b@localhost ~/Cloud
hadoop jar /home/btech/22/sridhar.tuli22b/Cloud/hadoop-3.4.0/share/hadoop/tools/lib/hadoop-streaming-3.4.0.jar \
>     -input /home/btech/22/sridhar.tuli22b/Cloud/big.txt \
>     -output /home/btech/22/sridhar.tuli22b/outputdir \
>     -mapper /home/btech/22/sridhar.tuli22b/Cloud/mapper.py \
>     -combiner /home/btech/22/sridhar.tuli22b/Cloud/combinerUniqueWords.py \
>     -reducer /home/btech/22/sridhar.tuli22b/Cloud/reducer.py
2024-11-08 11:26:55,232 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-11-08 11:26:55,346 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2024-11-08 11:26:55,347 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2024-11-08 11:26:55,365 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
```



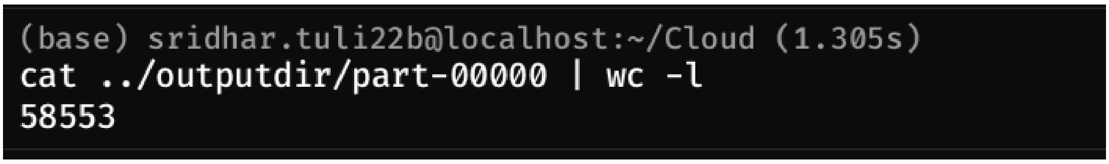
```
base sridhar.tuli22b@localhost ~/Cloud (7.761s)
hadoop jar /home/btech/22/sridhar.tuli22b/Cloud/hadoop-3.4.0/share/hadoop/tools/lib/hadoop-streaming-3.4.0.jar \
>     -input /home/btech/22/sridhar.tuli22b/Cloud/big.txt \
>     -output /home/btech/22/sridhar.tuli22b/outputdir \
>     -mapper /home/btech/22/sridhar.tuli22b/Cloud/mapper.py \
>     -combiner /home/btech/22/sridhar.tuli22b/Cloud/combinerUniqueWords.py \
>     -reducer /home/btech/22/sridhar.tuli22b/Cloud/reducer.py
        Map-Reduce Framework
                Map input records=128457
                Map output records=1095695
                Map output bytes=9706466
                Map output materialized bytes=853564
                Input split bytes=101
                Combine input records=1095695
                Combine output records=58553
                Reduce input groups=58553
                Reduce shuffle bytes=853564
                Reduce input records=58553
                Reduce output records=58553
                Spilled Records=117106
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=14
                Total committed heap usage (bytes)=1564475392
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=6488666
        File Output Format Counters
                Bytes Written=742216
2024-11-08 11:27:01,052 INFO streaming.StreamJob: Output directory: /home/btech/22/sridhar.tuli22b/outputdir
```

OUTPUT



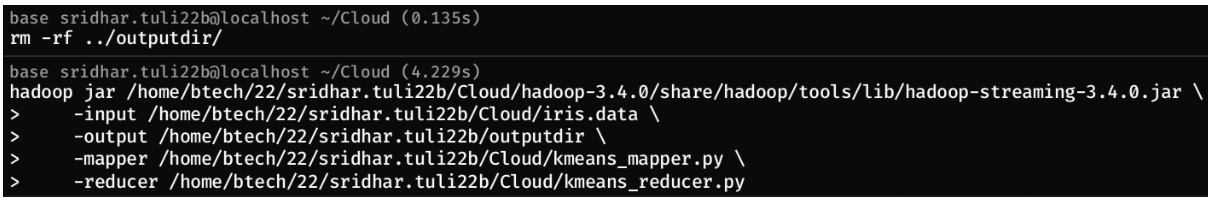
```
base sridhar.tuli22b@localhost ~/Cloud (5.708s)
cat ../outputdir/part-00000
|illinois,1
|indiana,1
|iowa,1
|kansas,1
|kentucky,1
|louisiana,1
|maine,1
|maryland,1
|massachusetts,1
|michigan,1
|minnesota,1
|mississippi,1
|missouri,1
|montana,1
|nebraska,1
|nevada,1
|new,4
|north,2
|not,1
|ohio,1
|oklahoma,1
|oregon,1
|pennsylvania,1
|rhode,1
|south,2
|tennessee,1
|texas,1
|united,1
|utah,1
|vermont,1
|virginia,1
|washington,1
|west,1
|wisconsin,1
|wyoming,1
```

LINE COUNT



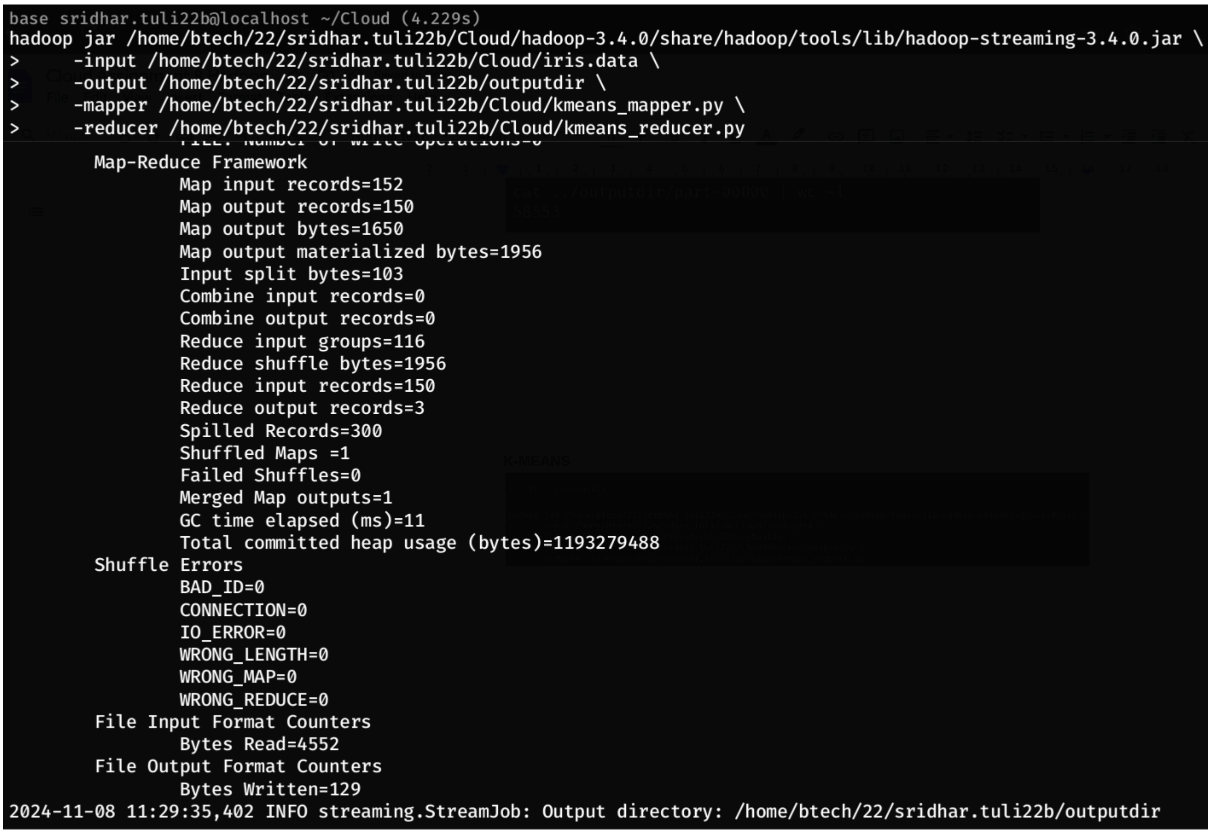
```
(base) sridhar.tuli22b@localhost:~/Cloud (1.305s)
cat ../outputdir/part-00000 | wc -l
58553
```

K-MEANS



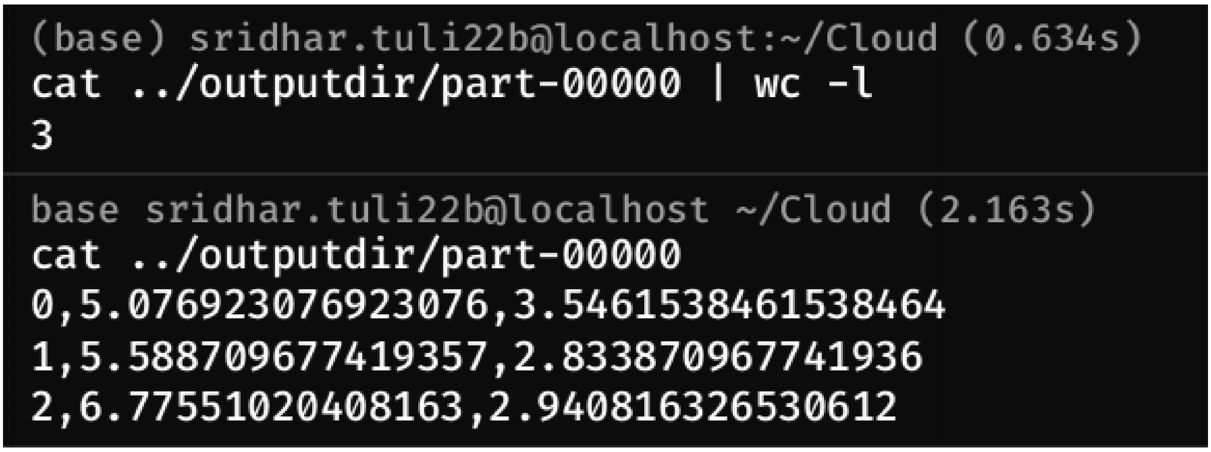
```
base sridhar.tuli22b@localhost ~/Cloud (0.135s)
rm -rf ../outputdir/

base sridhar.tuli22b@localhost ~/Cloud (4.229s)
hadoop jar /home/btech/22/sridhar.tuli22b/Cloud/hadoop-3.4.0/share/hadoop/tools/lib/hadoop-streaming-3.4.0.jar \
>     -input /home/btech/22/sridhar.tuli22b/Cloud/iris.data \
>     -output /home/btech/22/sridhar.tuli22b/outputdir \
>     -mapper /home/btech/22/sridhar.tuli22b/Cloud/kmeans_mapper.py \
>     -reducer /home/btech/22/sridhar.tuli22b/Cloud/kmeans_reducer.py
```

OUTPUT



```
base sridhar.tuli22b@localhost ~/Cloud (4.229s)
hadoop jar /home/btech/22/sridhar.tuli22b/Cloud/hadoop-3.4.0/share/hadoop/tools/lib/hadoop-streaming-3.4.0.jar \
>     -input /home/btech/22/sridhar.tuli22b/Cloud/iris.data \
>     -output /home/btech/22/sridhar.tuli22b/outputdir \
>     -mapper /home/btech/22/sridhar.tuli22b/Cloud/kmeans_mapper.py \
>     -reducer /home/btech/22/sridhar.tuli22b/Cloud/kmeans_reducer.py
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=152
                Map output records=150
                Map output bytes=1650
                Map output materialized bytes=1956
                Input split bytes=103
                Combine input records=0
                Combine output records=0
                Reduce input groups=116
                Reduce shuffle bytes=1956
                Reduce input records=150
                Reduce output records=3
                Spilled Records=300
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=11
                Total committed heap usage (bytes)=1193279488
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=4552
        File Output Format Counters
                Bytes Written=129
2024-11-08 11:29:35,402 INFO streaming.StreamJob: Output directory: /home/btech/22/sridhar.tuli22b/outputdir
```

AFTER CLASSIFICATION



```
(base) sridhar.tuli22b@localhost:~/Cloud (0.634s)
cat ../outputdir/part-00000 | wc -l
3

base sridhar.tuli22b@localhost ~/Cloud (2.163s)
cat ../outputdir/part-00000
0,5.076923076923076,3.5461538461538464
1,5.588709677419357,2.833870967741936
2,6.77551020408163,2.940816326530612
```

Cluster 0: The centroid of cluster 0 is at coordinates (5.0769, 3.5462).
Cluster 1: The centroid of cluster 1 is at coordinates (5.5887, 2.8339).
Cluster 2: The centroid of cluster 2 is at coordinates (6.7755, 2.9408).