CSC 555: Final

Name: Nachiketh Reddy
ID: 2117731
I have done this exam entirely on my own. I have not consulted with friends or consulted online resources that have the solution to the exam.

QUESTION 1:

Clustering Write Python code within the four files provided. Each file has instructions to write code as described in the Python triple quotes, i.e., "'TODO:...'". You must write code using these files and not write your own files. To write code you will need to be familiar with numpy. The tutorial at this site https: //numpy.org/doc/stable/user/quickstart.html will be sufficient. Once you have generated data, generated the centers, and written the code in all four files, test it with: cat kmeans data.csv | python3 mapper-kmeans.py | sort -n | python3 reducer-kmeans.py > centers.txt Now repeat the process and run 4 iterations of kMeans. Remember to change the centers file in mapper-kmeans.py everytime to correspond to the centers generated in the previous iteration. Note, you can maintain centers as centers[i].txt where [i] is the centers output by the i th iteration. The first centers file must be generated apriori and is used by mapper-kmeans.py Submit the following:
• The initial centers.txt and the kmeans data.csv. Please zip the kmeans data.csv. • All four Python files. • The final centers.txt after 4 iterations. • The Hadoop command used and a screenshot of running on Hadoop cluster with time information for each iteration.
Initial centers.txt:

```
$ cat centers.txt
0.418506478974507,0.08759190798437044,0.6202098841228467,0.3340070756262328,0.7613721028137473
0.798336886574648,0.4736009866774325 6,0.12065307097839384,0.5759179675004569,0.4655673944438966
0.18810268101729843,0.9971773467503174,0.5126709983572116,0.9026866448432411,0.3620564522378126
0.3562429889627011,0.039461805404032546,0.10208725537669039,0.9866256772840983,0.45061509541729616
0.13902698083927612,0.9338942588983641,0.022573079076542046,0.36246402492945984,0.07085101418126205
[ec2-user@ip-172-31-62-135 ~] 2024-03-22 16:56:07
$
```

i-062d65a3dd2e0d019 (nachiketh_server)
PublicIPs: 34.207.253.48 PrivateIPs: 172.31.62.135

First Run:
time hadoop jar hadoop-streaming-2.6.4.jar -input /data/kmeans_data.csv -output /data/out -mapper mapper-kmeans.py -reducer reducer-kmeans.py -file mapper-kmeans.py -file reducer-kmeans.py -file centers.txt

```
[ec2-user@ip-172-31-62-135 container_1711125434068_0003_01_000005] 2024-03-22 16:48:26
$ cd ~
[ec2-user@ip-172-31-62-135 ~] 2024-03-22 16:49:01
$ time hadoop jar hadoop-streaming-2.6.4.jar -input /data/kmeans_data.csv -output /data/out -mapper mapper-kmeans.py -reducer reducer-kmeans.py -file mapper-kmeans.py -file reducer-kmeans.py -file centers.txt
24/03/22 16:49:28 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [mapper-kmeans.py, reducer-kmeans.py, centers.txt, /tmp/hadoop-unjar2284059276364153829/] [] /tmp/streamjob1207521121587065422.jar tmpDir=null
24/03/22 16:49:28 INFO client.RMProxy: Connecting to ResourceManager at /172.31.62.135:8032
24/03/22 16:49:28 INFO client.RMProxy: Connecting to ResourceManager at /172.31.62.135:8032
24/03/22 16:49:29 INFO mapred.FileInputFormat: Total input paths to process : 1
24/03/22 16:49:29 INFO net.NetworkTopology: Adding a new node: /default-rack/172.31.62.135:50010
24/03/22 16:49:29 INFO mapreduce.JobSubmitter: number of splits:2
24/03/22 16:49:29 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1711125434068_0004
24/03/22 16:49:29 INFO impl.YarnClientImpl: Submitted application application_1711125434068_0004
24/03/22 16:49:29 INFO mapreduce.Job: The url to track the job: http://ip-172-31-62-135.ec2.internal:8088/proxy/application_1711125434068_0004/
24/03/22 16:49:29 INFO mapreduce.Job: Running job: job_1711125434068_0004
24/03/22 16:49:34 INFO mapreduce.Job: Job job_1711125434068_0004 running in uber mode : false
24/03/22 16:49:34 INFO mapreduce.Job:  map 0% reduce 0%
24/03/22 16:49:44 INFO mapreduce.Job:  map 6% reduce 0%
24/03/22 16:49:47 INFO mapreduce.Job:  map 9% reduce 0%
24/03/22 16:49:50 INFO mapreduce.Job:  map 13% reduce 0%
24/03/22 16:49:53 INFO mapreduce.Job:  map 16% reduce 0%
24/03/22 16:49:56 INFO mapreduce.Job:  map 19% reduce 0%
24/03/22 16:49:59 INFO mapreduce.Job:  map 23% reduce 0%
24/03/22 16:50:02 INFO mapreduce.Job:  map 26% reduce 0%
24/03/22 16:50:05 INFO mapreduce.Job:  map 29% reduce 0%
24/03/22 16:50:08 INFO mapreduce.Job:  map 32% reduce 0%
24/03/22 16:50:10 INFO mapreduce.Job:  map 50% reduce 0%
24/03/22 16:50:21 INFO mapreduce.Job:  map 56% reduce 0%
24/03/22 16:50:24 INFO mapreduce.Job:  map 59% reduce 0%
24/03/22 16:50:27 INFO mapreduce.Job:  map 62% reduce 0%
24/03/22 16:50:30 INFO mapreduce.Job:  map 65% reduce 0%
24/03/22 16:50:33 INFO mapreduce.Job:  map 69% reduce 0%
24/03/22 16:50:36 INFO mapreduce.Job:  map 72% reduce 0%
24/03/22 16:50:39 INFO mapreduce.Job:  map 75% reduce 0%
24/03/22 16:50:42 INFO mapreduce.Job:  map 78% reduce 0%
24/03/22 16:50:45 INFO mapreduce.Job:  map 81% reduce 0%
24/03/22 16:50:48 INFO mapreduce.Job:  map 100% reduce 0%
24/03/22 16:50:57 INFO mapreduce.Job:  map 100% reduce 81%
24/03/22 16:51:00 INFO mapreduce.Job:  map 100% reduce 89%
24/03/22 16:51:03 INFO mapreduce.Job:  map 100% reduce 97%
24/03/22 16:51:05 INFO mapreduce.Job:  map 100% reduce 100%
24/03/22 16:51:05 INFO mapreduce.Job: Job job_1711125434068_0004 completed successfully
24/03/22 16:51:05 INFO mapreduce.Job: Counters: 50
```

i-062d65a3dd2e0d019 (nachiketh_server)
PublicIPs: 34.207.253.48 PrivateIPs: 172.31.62.135

```
aws   Services   Search                                   [Alt+S]
                Total time spent by all map tasks (ms)=70312
                Total time spent by all reduce tasks (ms)=15344
                Total vcore-milliseconds taken by all map tasks=70312
                Total vcore-milliseconds taken by all reduce tasks=15344
                Total megabyte-milliseconds taken by all map tasks=421872000
                Total megabyte-milliseconds taken by all reduce tasks=92064000
        Map-Reduce Framework
                Map input records=2000000
                Map output records=2000000
                Map output bytes=200697359
                Map output materialized bytes=204697371
                Input split bytes=186
                Combine input records=0
                Combine output records=0
                Reduce input groups=2000000
                Reduce shuffle bytes=204697371
                Reduce input records=2000000
                Reduce output records=5
                Spilled Records=6000000
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=199
                CPU time spent (ms)=86850
                Physical memory (bytes) snapshot=857776128
                Virtual memory (bytes) snapshot=9967841280
                Total committed heap usage (bytes)=838336512
        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=192701455
        File Output Format Counters
                Bytes Written=480
24/03/22 16:51:05 INFO streaming.StreamJob: Output directory: /data/out

real    1m38.293s
user    0m3.957s
sys     0m0.246s
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 34.207.253.48 PrivateIPs: 172.31.62.135

centers1.txt

```
$ hadoop fs -cat /data/out/part-00000 | more
0.4466750888751059,0.3252244435386771,0.6633358814501816,0.3699391537086328,0.6190676306814223
0.7520720779788497,0.5478180588334175,0.3475748212662098,0.503967543531433,0.4704351345767741
0.3426654567922555,0.7763893876610183,0.6377502557562773,0.7037533394701059,0.45552921303595967
0.32200687462807565,0.2042678290089987,0.3010958777611171,0.809901256192348,0.427933120977061
0.25406158464323236,0.7309643097996597,0.280721980197176,0.27361726142219434,0.2685530185949394
[ec2-user@ip-172-31-62-135 ~] 2024-03-22 17:01:06
$
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 34.207.253.48 PrivateIPs: 172.31.62.135

Second Run:

time hadoop jar hadoop-streaming-2.6.4.jar -input /data/kmeans_data.csv -output /data/out_1 -mapper mapper-kmeans.py -reducer reducer-kmeans.py -file mapper-kmeans.py -file reducer-kmeans.py -file centers1.txt

```
24/03/22 17:06:28 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=409394754
                FILE: Number of bytes written=614424194
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=192701641
                HDFS: Number of bytes written=484
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=455628
                Total time spent by all reduces in occupied slots (ms)=91866
                Total time spent by all map tasks (ms)=75938
                Total time spent by all reduce tasks (ms)=15311
                Total vcore-milliseconds taken by all map tasks=75938
                Total vcore-milliseconds taken by all reduce tasks=15311
                Total megabyte-milliseconds taken by all map tasks=455628000
                Total megabyte-milliseconds taken by all reduce tasks=91866000
        Map-Reduce Framework
                Map input records=2000000
                Map output records=2000000
                Map output bytes=200697359
                Map output materialized bytes=204697371
                Input split bytes=186
                Combine input records=0
                Combine output records=0
                Reduce input groups=2000000
                Reduce shuffle bytes=204697371
                Reduce input records=2000000
                Reduce output records=5
                Spilled Records=6000000
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=227
                CPU time spent (ms)=92940
                Physical memory (bytes) snapshot=839688192
                Virtual memory (bytes) snapshot=9970372608
                Total committed heap usage (bytes)=789577728
        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=192701455
        File Output Format Counters
                Bytes Written=484
24/03/22 17:06:28 INFO streaming.StreamJob: Output directory: /data/out_1

real    1m44.789s
user    0m3.902s
sys     0m0.366s
[ec2-user@ip-172-31-62-135 ~] 2024-03-22 17:06:28
$
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 34.207.253.48 PrivateIPs: 172.31.62.135

centers2.txt:

```
[ec2-user@ip-172-31-62-135 ~] 2024-03-22 17:06:28
$ hadoop fs -cat /data/out_1/part-00000 | more
0.4734409778155092,0.3143391508162799 4,0.6986501515506048,0.33940488359268,0.6265724075484513
0.8031124396264666,0.5546227727457731,0.3595995575618838,0.49196922450537167,0.4901047027459025
0.3831126043898117,0.761100105539219,0.6763279032276892,0.7093220751550083,0.4929102468820437
0.3487814135022848 6,0.2332997342968309 6,0.3339663188372609,0.7623973942110422,0.4473854823347234
0.2818982942644900 3,0.6805021345177569,0.3121872970906006 7,0.2586042314509457 4,0.3326422763007367
[ec2-user@ip-172-31-62-135 ~] 2024-03-22 17:07:29
$
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 34.207.253.48 PrivateIPs: 172.31.62.135

Third Run:

time hadoop jar hadoop-streaming-2.6.4.jar -input /data/kmeans_data.csv -output /data/out_2 -mapper mapper-kmeans.py -reducer reducer-kmeans.py -file mapper-kmeans.py -file reducer-kmeans.py -file centers2.txt

```
24/03/22 17:11:02 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=409394754
                FILE: Number of bytes written=614424194
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=192701641
                HDFS: Number of bytes written=486
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=434508
                Total time spent by all reduces in occupied slots (ms)=91596
                Total time spent by all map tasks (ms)=72418
                Total time spent by all reduce tasks (ms)=15266
                Total vcore-milliseconds taken by all map tasks=72418
                Total vcore-milliseconds taken by all reduce tasks=15266
                Total megabyte-milliseconds taken by all map tasks=434508000
                Total megabyte-milliseconds taken by all reduce tasks=91596000
        Map-Reduce Framework
                Map input records=2000000
                Map output records=2000000
                Map output bytes=200697359
                Map output materialized bytes=204697371
                Input split bytes=186
                Combine input records=0
                Combine output records=0
                Reduce input groups=2000000
                Reduce shuffle bytes=204697371
                Reduce input records=2000000
                Reduce output records=5
                Spilled Records=6000000
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=211
                CPU time spent (ms)=90180
                Physical memory (bytes) snapshot=838172672
                Virtual memory (bytes) snapshot=9972330496
                Total committed heap usage (bytes)=804782080
        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=192701455
        File Output Format Counters
                Bytes Written=486
24/03/22 17:11:02 INFO streaming.StreamJob: Output directory: /data/out_2

real    1m40.371s
user    0m4.009s
sys     0m0.349s
[ec2-user@ip-172-31-62-135 ~] 2024-03-22 17:11:02
$
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 34.207.253.48 PrivateIPs: 172.31.62.135

Centers3.txt:

```
$ hadoop fs -cat /data/out_2/part-00000 | more
0.4937146240474645,0.3036055110369187,0.7221833498730471,0.3211809132270863 6,0.6218470497306814
0.8201501479677525,0.5599202231741127,0.3580808876313664,0.4863099622151565 4,0.4960995401869649 5
0.4076023877526470 7,0.7521683493085846,0.6931897494272781,0.7136434667723321,0.5082152874760778
0.3634028139421850 4,0.2444766586713419 7,0.3484375930606069,0.746892746410468,0.457046003110177
0.2912114204948252 3,0.6605776564436759,0.3258643024647322,0.2568438608784413,0.3676739830340884 4
[ec2-user@ip-172-31-62-135 ~] 2024-03-22 17:12:47
$
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 34.207.253.48 PrivateIPs: 172.31.62.135

Fourth run:

time hadoop jar hadoop-streaming-2.6.4.jar -input /data/kmeans_data.csv -output /data/out_3 -mapper mapper-kmeans.py -reducer reducer-kmeans.py -file mapper-kmeans.py -file reducer-kmeans.py -file centers3.txt

```
24/03/22 17:15:49 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=409394754
                FILE: Number of bytes written=614424194
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=192701641
                HDFS: Number of bytes written=485
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=453996
                Total time spent by all reduces in occupied slots (ms)=93816
                Total time spent by all map tasks (ms)=75666
                Total time spent by all reduce tasks (ms)=15636
                Total vcore-milliseconds taken by all map tasks=75666
                Total vcore-milliseconds taken by all reduce tasks=15636
                Total megabyte-milliseconds taken by all map tasks=453996000
                Total megabyte-milliseconds taken by all reduce tasks=93816000
        Map-Reduce Framework
                Map input records=2000000
                Map output records=2000000
                Map output bytes=200697359
                Map output materialized bytes=204697371
                Input split bytes=186
                Combine input records=0
                Combine output records=0
                Reduce input groups=2000000
                Reduce shuffle bytes=204697371
                Reduce input records=2000000
                Reduce output records=5
                Spilled Records=6000000
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=213
                CPU time spent (ms)=93230
                Physical memory (bytes) snapshot=837906432
                Virtual memory (bytes) snapshot=9970728960
                Total committed heap usage (bytes)=760217600
        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=192701455
        File Output Format Counters
                Bytes Written=485
24/03/22 17:15:49 INFO streaming.StreamJob: Output directory: /data/out_3

real    1m45.476s
user    0m4.127s
sys     0m0.303s
[ec2-user@ip-172-31-62-135 ~] 2024-03-22 17:15:49
$
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 34.207.253.48 PrivateIPs: 172.31.62.135

Centers4.txt:

```
[ec2-user@ip-172-31-62-135 ~] 2024-03-22 17:15:49
$ hadoop fs -cat /data/out_3/part-00000 | more
0.5103926053310995,0.2977545901367192,0.7366440296430301,0.31258292999415127,0.6101416282578812
0.8242878188037962,0.5655850345644502,0.3483211768832058,0.48126983360050707,0.497323176535033
0.42362716273242834,0.7463217670226793,0.7030348101671381,0.715393510079285,0.5143133002628412
0.37082399732522636,0.24839848545648657,0.3558784053879261 4,0.7425507539694319,0.4630451944577867
0.29027488652129885,0.6523560172490893,0.331962013239862,0.2576935771924344,0.39065532819201376
[ec2-user@ip-172-31-62-135 ~] 2024-03-22 17:16:38
$
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 34.207.253.48 PrivateIPs: 172.31.62.135

QUESTION 2:

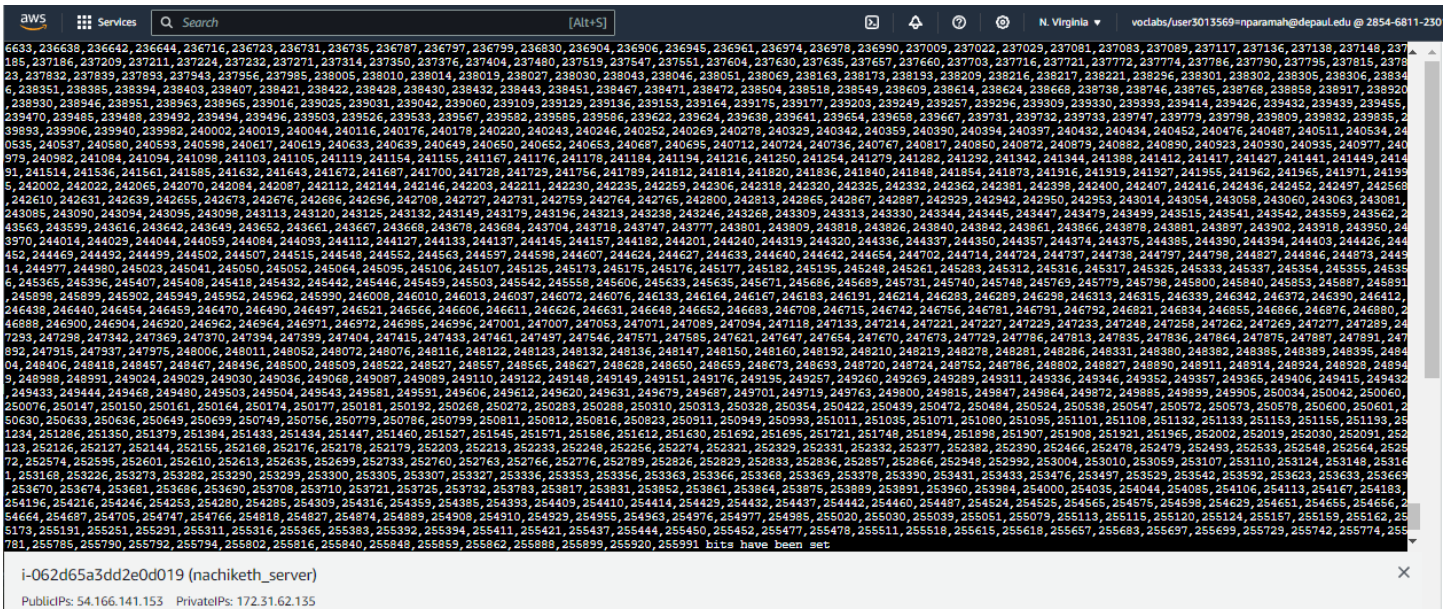
BloomFilter-based 2-way map-side Join The objective of this question is to implement a 1-pass MR that uses bloom filter for the larger table. Write Python code within the three files provided. Each file has instructions to write code as described in the Python triple quotes, i.e., "'TODO:..."'. You must write code using these files and not write your own files. Consider the following query:

select * from lineorder, dwdate where lo_orderdate = d_datekey and d_sellingseason = 'Fall'

Lineorder and Dwdate are to be downloaded from [http://cdmgcsarprd01.dpu.depaul.edu/ CSC555/SSBM1/dwdate.tbl](http://cdmgcsarprd01.dpu.depaul.edu/CSC555/SSBM1/dwdate.tbl) http://cdmgcsarprd01.dpu.depaul.edu/CSC555/SSBM1/lineorder.tbl

In the map phase, the bloom filter of the large table (lineorder) will be read. The small table (dwdate) will also be read. In the map phase, make both the where clause checks and output the join result. The reduce phase just passes the result. The bloomfilter should be setup as described in bloomfilter.py.

Python bloomfilter.py: bits set



time hadoop jar hadoop-streaming-2.6.4.jar -input /data/dwdate.tbl -output /data/output1 -mapper BFmapper.py -reducer BFreducer.py -file BFmapper.py -file BFreducer.py -file bloom_filter.bloom

```
aws  Services  Search [Alt+S]
[ec2-user@ip-172-31-62-135 ~] 2024-03-22 16:38:18
$ time hadoop jar hadoop-streaming-2.6.4.jar -input /data/dwdate.tbl -output /data/output1 -mapper BFmapper.py -reducer BFreducer.py -file BFmapper.py -file BFreducer.py -file bloom_filter.bloom
24/03/22 16:38:26 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [BFmapper.py, BFreducer.py, bloom_filter.bloom, /tmp/hadoop-unjar2103746955254485 56/] [] /tmp/streamjob6254127611655366535.jar tmpDir=null
24/03/22 16:38:27 INFO client.RMProxy: Connecting to ResourceManager at /172.31.62.135:8032
24/03/22 16:38:27 INFO client.RMProxy: Connecting to ResourceManager at /172.31.62.135:8032
24/03/22 16:38:28 INFO mapred.FileInputFormat: Total input paths to process : 1
24/03/22 16:38:28 INFO mapreduce.JobSubmitter: number of splits:2
24/03/22 16:38:28 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1711125434068_0002
24/03/22 16:38:28 INFO impl.YarnClientImpl: Submitted application application_1711125434068_0002
24/03/22 16:38:28 INFO mapreduce.Job: The url to track the job: http://ip-172-31-62-135.ec2.internal:8088/proxy/application_1711125434068_0002/
24/03/22 16:38:28 INFO mapreduce.Job: Running job: job_1711125434068_0002
24/03/22 16:38:34 INFO mapreduce.Job: Job job_1711125434068_0002 running in uber mode : false
24/03/22 16:38:34 INFO mapreduce.Job:  map 0% reduce 0%
24/03/22 16:38:40 INFO mapreduce.Job:  map 50% reduce 0%
24/03/22 16:38:44 INFO mapreduce.Job:  map 100% reduce 0%
24/03/22 16:38:48 INFO mapreduce.Job:  map 100% reduce 100%
24/03/22 16:38:48 INFO mapreduce.Job: Job job_1711125434068_0002 completed successfully
24/03/22 16:38:48 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=9522
                FILE: Number of bytes written=351044
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=234237
                HDFS: Number of bytes written=8784
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=29052
                Total time spent by all reduces in occupied slots (ms)=14628
                Total time spent by all map tasks (ms)=4842
                Total time spent by all reduce tasks (ms)=2438
                Total vcore-milliseconds taken by all map tasks=4842
                Total vcore-milliseconds taken by all reduce tasks=2438
                Total megabyte-milliseconds taken by all map tasks=29052000
                Total megabyte-milliseconds taken by all reduce tasks=14628000
        Map-Reduce Framework
                Map input records=2556
                Map output records=366
                Map output bytes=8784
                Map output materialized bytes=9528
                Input split bytes=176
                Combine input records=0
                Combine output records=0
                Reduce input groups=366
i-062d65a3dd2e0d019 (nachiketh_server)
PublicIPs: 34.207.253.48  PrivateIPs: 172.31.62.135
```

Run Time and Output:

```
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=234061
        File Output Format Counters
                Bytes Written=8784
24/03/22 16:38:48 INFO streaming.StreamJob: Output directory: /data/output1

real    0m22.342s
user    0m3.758s
sys     0m0.194s
[ec2-user@ip-172-31-62-135 ~] 2024-03-22 16:38:48
$ hdfs dfs -cat /data/output1/part-00000 | more
19920901|19920901|Fall
19920902|19920902|Fall
19920903|19920903|Fall
19920904|19920904|Fall
19920905|19920905|Fall
19920906|19920906|Fall
19920907|19920907|Fall
19920908|19920908|Fall
19920909|19920909|Fall
19920910|19920910|Fall
19920911|19920911|Fall
19920912|19920912|Fall
19920913|19920913|Fall
19920914|19920914|Fall
19920915|19920915|Fall
19920916|19920916|Fall
19920917|19920917|Fall
19920918|19920918|Fall
19920919|19920919|Fall
19920920|19920920|Fall
19920921|19920921|Fall
19920922|19920922|Fall
19920923|19920923|Fall
19920924|19920924|Fall
19920925|19920925|Fall
19920926|19920926|Fall
19920927|19920927|Fall
19920928|19920928|Fall
19920929|19920929|Fall
19920930|19920930|Fall
19921001|19921001|Fall
19921002|19921002|Fall
19921003|19921003|Fall
19921004|19921004|Fall
19921005|19921005|Fall
19921006|19921006|Fall
19921007|19921007|Fall
19921008|19921008|Fall
```

i-062d65a3dd2e0d019 (nachiketh_server)

PublicIPs: 34.207.253.48 PrivateIPs: 172.31.62.135