

Engineering Cloud Computing
Assignment 1
Arpita Ajit Welling (aawellin@iu.edu)

Q1:

For question 1, we were given an access log and we had to find the top 3 ip addresses by count for each hour.

To do this, I started by using the mapreduce code that was provided. First, I tried to get the ip counts in sorted order. The mapper code for this problem would just assign 1 count to each and every entry in the input. The reducer code will then actually count the number of input for that particular ip. Reducer will also sort this count. I have used a dictionary to maintain the count of each ip.

The problem focuses on finding the ip counts according to hour granularity and then return top 3 ip addresses for each hour. For this, I have used 2 MapReduce. The first mapper would just assign each hour-ip address pair a count of 1 and get all possible inputs. The first reducer will count the number of different hour and ip-addresses combinations and give it's count. The second mapper will take the first reducer's output as input and count the ip addresses by hour granularity. The second mapper will also give top-3 ip addresses of each hour locally. The second reducer will do the same thing but give the top-3 ip addresses by hour globally.

To chain the 2 mapreduce, I have used a bash script (hadoop_MR1.sh) to manually send the output of first mapreduce to the second mapreduce.

I also wrote another bash script (hadoop_remove.sh) that deletes the output files later.

One of the challenges here was to understand how the reducer tasks work and exactly what input is the reducer getting. To solve this problem, I also ran the mapper code without hadoop and saw what was the output that I was getting before running the hadoop commands. I did the same with the reducer. As an experiment I also tried increasing the number of reduce tasks.

Below are the screenshots after running the bash file. We can see the output with hour granularity.

```

~/ECC_A1_code -- -zsh                                         /opt/homebrew/Cellar/hadoop/3.3.4/libexec/etc/hadoop -- ssh localhost
(base) arpitawelling@arpitas-MacBook-Pro ~ % bash hadoop MRJob
2022-11-06 19:44:40.233 WARN streaming.StreamJob: File option is deprecated, please use generic option -files instead.
2022-11-06 19:44:40.233 WARN streaming.StreamJob: File option is deprecated, please use generic option -files instead.
packageJobJar: [/Users/arpitawelling/ECC_A1_code/mapper_time.py, /Users/arpitawelling/ECC_A1_code/reducer_ip.py] [] /var/folders/35/j_3430ps66ldz?zwycq4_fdr=000gn/T/streamjob3521137588783849426.jar tmpDir
=null
2022-11-06 19:44:40.716 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2022-11-06 19:44:40.758 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 seconds.
2022-11-06 19:44:40.760 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2022-11-06 19:44:40.760 INFO impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2022-11-06 19:44:40.894 INFO mapred.FileOutputFormat: Total input file(s): process-1
2022-11-06 19:44:40.968 INFO mapreduce.JobSubmitter: number of splits:26
2022-11-06 19:44:41.047 INFO mapreduce.JobSubmitter: Adding a new node: /default-rack/127.0.0.1:9866
2022-11-06 19:44:40.968 INFO mapreduce.JobSubmitter: number of splits:26
2022-11-06 19:44:41.047 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1985950695_0001
2022-11-06 19:44:41.047 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-11-06 19:44:41.047 INFO mapred.LocalDistributedCacheManager: Localized file:/Users/arpitawelling/ECC_A1_code/mapper_time.py as file:/opt/homebrew/Cellar/hadoop/tmp/mapred/local/job_local1985950695_0001_0001_d041180a-143b-43f4-aefc-4e0a2a2a2a2a
2022-11-06 19:44:41.050 INFO mapred.LocalDistributedCacheManager: Localized file:/Users/arpitawelling/ECC_A1_code/reducer_ip.py as file:/opt/homebrew/Cellar/hadoop/tmp/mapred/local/job_local1985950695_0001_000665750-1bc2-43a0-b2bd-95aaeddb597/reducer_ip.py
2022-11-06 19:44:41.182 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2022-11-06 19:44:41.183 INFO mapreduce.Job: Job Running job: job_local1985950695_0001
2022-11-06 19:44:41.184 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2022-11-06 19:44:41.184 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
2022-11-06 19:44:41.185 INFO mapred.LocalJobRunner: OutputCommitter Algorithm version is 2
2022-11-06 19:44:41.188 INFO mapred.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2022-11-06 19:44:41.189 INFO mapred.LocalJobRunner: Waiting for map tasks
2022-11-06 19:44:41.204 INFO mapred.LocalJobRunner: Starting task: attempt_local1985950695_0001_m_000000_0
2022-11-06 19:44:41.204 INFO mapred.FileOutputCommitter: File Output Committer Algorithm version is 2
2022-11-06 19:44:41.204 INFO mapred.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2022-11-06 19:44:41.289 INFO mapred.ProcfsBasedProcessTree: currently is supported only on Linux.
2022-11-06 19:44:41.289 INFO mapred.Task: Using ProcfsBasedProcessTree
2022-11-06 19:44:41.392 INFO mapred.PipeMapTask: Processing split hdfs://localhost:8020/user/access.log:3355443200+146997623
2022-11-06 19:44:41.371 INFO mapred.MapTask: numReduceTasks: 1
2022-11-06 19:44:41.388 INFO mapred.MapTask: (EQUATOR) 0 kv1 26214396((194857584)
2022-11-06 19:44:41.388 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2022-11-06 19:44:41.388 INFO mapred.MapTask: soft limit at 83868000
2022-11-06 19:44:41.388 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2022-11-06 19:44:41.388 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2022-11-06 19:44:41.392 INFO mapred.PipeMapRed exec [/Users/arpitawelling//mapper_time.py]
2022-11-06 19:44:41.394 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
2022-11-06 19:44:41.398 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
2022-11-06 19:44:41.398 INFO Configuration.deprecation: map:input.fil is deprecated. Instead, use mapreduce.map.input.file
2022-11-06 19:44:41.398 INFO Configuration.deprecation: map:input.length is deprecated. Instead, use mapreduce.map.input.length
2022-11-06 19:44:41.398 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
2022-11-06 19:44:41.398 INFO Configuration.deprecation: mapred.confs is deprecated. Instead, use mapreduce.map.confs
2022-11-06 19:44:41.398 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.map.skip.records
2022-11-06 19:44:41.398 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
2022-11-06 19:44:41.397 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
2022-11-06 19:44:41.397 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
2022-11-06 19:44:41.397 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
2022-11-06 19:44:41.397 INFO Configuration.deprecation: user.timezone is deprecated. Instead, use mapreduce.job.timezone
2022-11-06 19:44:41.400 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
2022-11-06 19:44:41.400 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
2022-11-06 19:44:41.452 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
2022-11-06 19:44:41.456 INFO streaming.PipeMapRed: Records R/W=1573/1
2022-11-06 19:44:41.492 INFO streaming.PipeMapRed: R/W/S=10000/8828/0 in:NA [rec/s] out:NA [rec/s]
2022-11-06 19:44:41.651 INFO streaming.PipeMapRed: R/W/S=10000/9906/0 in:NA [rec/s] out:NA [rec/s]
2022-11-06 19:44:41.651 INFO streaming.PipeMapRed: R/W/S=20000/0/0 in:NA [rec/s] out:NA [rec/s]
2022-11-06 19:44:41.943 INFO streaming.PipeMapRed: R/W/S=30000/0/0 in:NA [rec/s] out:NA [rec/s]
2022-11-06 19:44:42.965 INFO streaming.PipeMapRed: R/W/S=ErrorThread done

~/ECC_A1_code -- -zsh                                         /opt/homebrew/Cellar/hadoop/3.3.4/libexec/etc/hadoop -- ssh localhost
(base) arpitawelling@arpitas-MacBook-Pro ~ % 

```

```

~/ECC_A1_code -- -zsh          /opt/homebrew/Cellar/hadoop/3.3.4/libexec/etc/hadoop -- ssh localhost +
```

```
(base) arpitawelling@Ariptas-MacBook-Pro ~ % hdfs dfs -cat /user/ECC_A1-output.access_full/part-00000
2022-11-06 15:57:07,865 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-jar classes where applicable
[00:00]16.249.66.91 14298
[00:00]16.249.66.91 12232
[00:00]16.249.66.91 4291
[01:00]16.249.66.91 13874
[01:00]16.249.66.194 12485
[01:00]16.249.66.92 2924
[01:00]16.249.66.91 11097
[02:00]16.249.66.194 18945
[02:00]19.199.72.15 1448
[03:00]16.249.66.194 8644
[03:00]16.249.66.91 7914
[03:00]19.199.72.15 1275
[04:00]16.249.66.194 18895
[04:00]19.199.72.15 7971
[04:00]19.199.72.15 1551
[05:00]16.249.66.194 18534
[05:00]16.249.66.91 7835
[05:00]19.199.72.15 1921
[06:00]16.249.66.194 10283
[06:00]16.249.66.91 7968
[06:00]16.249.66.91 2851
[07:00]16.249.66.194 12257
[07:00]16.249.66.91 9116
[07:00]19.199.72.15 2295
[08:00]16.249.66.194 12964
[08:00]16.249.66.91 10237
[08:00]151.239.241.163 6256
[09:00]16.249.66.91 14833
[09:00]16.249.66.91 11450
[09:00]151.239.241.163 9169
[10:00]16.249.66.194 17292
[10:00]16.249.66.91 13213
[10:00]151.239.241.163 9824
[11:00]16.249.66.194 15572
[11:00]16.249.66.91 13631
[11:00]151.239.241.163 9142
[12:00]16.249.66.194 16966
[12:00]16.249.66.91 12656
[12:00]151.239.241.163 8564
[13:00]16.249.66.194 18372
[13:00]16.249.66.91 16166
[13:00]151.239.241.163 7881
[14:00]16.249.66.194 17249
[14:00]16.249.66.91 17893
[14:00]151.239.241.163 8786
[15:00]16.249.66.194 18273
[15:00]16.249.66.91 16662
[15:00]151.239.241.163 6558
[16:00]16.249.66.91 17849
[16:00]16.249.66.194 17172
[16:00]151.239.241.163 1487
[17:00]16.249.66.194 18954
[17:00]16.249.66.91 17107
[17:00]151.239.241.163 8571
[18:00]16.249.66.194 17531
[18:00]16.249.66.91 16727
[18:00]184.222.32.91 7157
[19:00]16.249.66.91 18911
[19:00]16.249.66.194 18549
[19:00]184.222.32.91 9876
[20:00]16.249.66.91 15834
[20:00]16.249.66.194 15729
[20:00]16.249.66.92 5589
[21:00]16.249.66.194 14875
[21:00]16.249.66.91 13783
[21:00]16.249.66.92 4552
[22:00]16.249.66.91 4494
[22:00]16.249.66.194 13576
[22:00]16.249.66.92 4981
[23:00]16.249.66.194 14355
[23:00]16.249.66.91 10902
[23:00]16.249.66.92 4259
```

```

~/ECC_A1_code -- -zsh          /opt/homebrew/Cellar/hadoop/3.3.4/libexec/etc/hadoop -- ssh localhost +
```

```
[04:00]16.249.66.194 10895
[04:00]16.249.66.91 7571
[04:00]16.249.66.194 1551
[05:00]16.249.66.194 18534
[05:00]16.249.66.91 7835
[05:00]19.199.72.15 1921
[06:00]16.249.66.194 10283
[06:00]16.249.66.91 7968
[06:00]16.249.66.91 2851
[07:00]16.249.66.194 12247
[07:00]16.249.66.91 11450
[07:00]19.199.72.15 2295
[08:00]16.249.66.194 12964
[08:00]16.249.66.91 10237
[08:00]151.239.241.163 6256
[09:00]16.249.66.194 14833
[09:00]16.249.66.91 11450
[09:00]151.239.241.163 9169
[10:00]16.249.66.194 17292
[10:00]16.249.66.91 13213
[10:00]151.239.241.163 9824
[11:00]16.249.66.194 15572
[11:00]16.249.66.91 13631
[11:00]151.239.241.163 8642
[12:00]16.249.66.194 17246
[12:00]16.249.66.91 12656
[12:00]151.239.241.163 8564
[13:00]16.249.66.194 18372
[13:00]16.249.66.91 16166
[13:00]151.239.241.163 7881
[14:00]16.249.66.194 17249
[14:00]16.249.66.91 17893
[14:00]151.239.241.163 8786
[15:00]16.249.66.194 18273
[15:00]16.249.66.91 16662
[15:00]151.239.241.163 6558
[16:00]16.249.66.91 17849
[16:00]16.249.66.194 17512
[16:00]16.249.66.91 12253
[17:00]16.249.66.194 1787
[17:00]16.249.66.91 15554
[17:00]16.249.66.91 17197
[17:00]151.239.241.163 8571
[18:00]16.249.66.194 17531
[18:00]16.249.66.91 16727
[18:00]184.222.32.91 7157
[19:00]16.249.66.91 18911
[19:00]16.249.66.194 18549
[19:00]184.222.32.91 9876
[20:00]16.249.66.91 15834
[20:00]16.249.66.194 15729
[20:00]16.249.66.92 5589
[21:00]16.249.66.194 14875
[21:00]16.249.66.91 13783
[21:00]16.249.66.92 4552
[22:00]16.249.66.91 4494
[22:00]16.249.66.194 13576
[22:00]16.249.66.92 4981
[23:00]16.249.66.194 14355
[23:00]16.249.66.91 10902
[23:00]16.249.66.92 4259
```

```
(base) arpitawelling@Ariptas-MacBook-Pro ~ %
```

Q2:

For the second problem, we had to perform a database search on the output of the first problem. The code will take the first problem's output as input and also take another range (0-1, 7-21,etc) as input.

E.g: If the range is 7-21, the program should output the top 3 ip addresses from time 8:00 to 12:00 (7:00 is excluded).

To do this, I designed a map reduce where the mapper will take in the range as argument and also take the output of the first question as input. It will then filter according to the range and give top-3 ip addresses locally. The reducer will take these top-3 ip addresses given by each mapper and give the global top-3 ip addresses within that range.

Below are the screenshots of this problem:

```
-[ECC_A1_code --> zsh] [opt/homebrew/Cellar/hadoop/3.3.4/bin/bexec@elijah:hadoop - ssh localhost] +  
(base) arpitawelling@Aritpas-MacBook-Pro ~ % /opt/homebrew/Cellar/hadoop/3.3.4/bin/hadoop jar /Users/rpitawelling/ECC_A1/hadoop-streaming*.jar -file /Users/rpitawelling/ECC_A1_code/mapper3.py -mapper3 -input /Users/rpitawelling/ECC_A1_code/reducer3.py -reducer /Users/rpitawelling/ECC_A1_code/reducer3.py -input /user/ECC_A1-output_access_full/part-00  
000  
2022-11-06 19:56:45.384 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.  
2022-11-06 19:56:45.384 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-Java classes where applicable  
packageJobJar: [/Users/rpitawelling/ECC_A1_code/mapper3.py, /Users/rpitawelling/ECC_A1_code/reducer3.py] [l] /var/folders/35/j_3430ps661dx9zwyco4_fdr000gn/T/streamjob6245822275099720504.jar tmpDir=null  
2022-11-06 19:56:48.929 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties  
2022-11-06 19:56:48.968 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 seconds.  
2022-11-06 19:56:48.968 INFO impl.MetricsSystemImpl: JobTracker metrics system started  
2022-11-06 19:56:49.303 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!  
2022-11-06 19:56:49.303 INFO mapred.JobSubmitter: Total number of tasks to process : 1  
2022-11-06 19:56:49.427 INFO mapreduce.JobSubmitter: Number of splits:1  
2022-11-06 19:56:49.215 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1936192088_0001  
2022-11-06 19:56:49.303 INFO mapred.LocalDistributedCacheManager: Localized file:/Users/rpitawelling/ECC_A1_code/mapper3.py as file:/opt/homebrew/Cellar/hadoop/hdfs/tmp/mapred/local/job_local1936192088_001_74c0d4358-6c7e-48d4-8efc-89d7fd874401/mapper3.py  
2022-11-06 19:56:49.319 INFO mapred.LocalDistributedCacheManager: Localized file:/Users/rpitawelling/ECC_A1_code/reducer3.py as file:/opt/homebrew/Cellar/hadoop/hdfs/tmp/mapred/local/job_local1936192088_001_74c0d4358-6c7e-48d4-8efc-89d7fd874401/reducer3.py  
000  
2022-11-06 19:56:49.342 INFO mapreduce.Job: The url to track the job: http://localhost:8080/  
2022-11-06 19:56:49.342 INFO mapred.LocalJobRunner: OutputCommitter set in config null  
2022-11-06 19:56:49.343 INFO mapred.LocalJobRunner: Job Running job: job_local1936192088_0001  
2022-11-06 19:56:49.344 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter  
2022-11-06 19:56:49.344 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2  
2022-11-06 19:56:49.344 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false  
2022-11-06 19:56:49.344 INFO mapred.LocalJobRunner: Starting task attempt_local1936192088_0001_m_000000_0  
2022-11-06 19:56:49.413 INFO output.FileOutputCommitter: File Output Committer Algorithm version is ?  
2022-11-06 19:56:49.413 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false  
2022-11-06 19:56:49.417 INFO util.ProcsBasedProcessTree: ProcsBasedProcessTree currently is supported only on Linux.  
2022-11-06 19:56:49.417 INFO mapred.Task: Using ResourceCalculatorProcessTree : null  
2022-11-06 19:56:49.419 INFO mapred.MapTask: Processing split: hdfs://localhost:8020/user/ECC_A1-output_access_full/part-00000:0+1892  
2022-11-06 19:56:49.450 INFO mapred.MapTask: numberOfRecordsTasks: 1  
2022-11-06 19:56:49.450 INFO mapred.MapTask: mapred.tip.id: 104057584  
2022-11-06 19:56:49.450 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100  
2022-11-06 19:56:49.450 INFO mapred.MapTask: soft limit at 83868698  
2022-11-06 19:56:49.450 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600  
2022-11-06 19:56:49.450 INFO mapred.MapTask: kvstart = 26214396; length = 6553600  
2022-11-06 19:56:49.452 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer  
2022-11-06 19:56:49.452 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/rpitawelling/.mappers3.py, -7-18]  
2022-11-06 19:56:49.452 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir  
2022-11-06 19:56:49.452 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir  
2022-11-06 19:56:49.458 INFO Configuration.deprecation: map-input.file is deprecated. Instead, use mapreduce.map.input.file  
2022-11-06 19:56:49.458 INFO Configuration.deprecation: map-input.length is deprecated. Instead, use mapreduce.map.input.length  
2022-11-06 19:56:49.459 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id  
2022-11-06 19:56:49.459 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition  
2022-11-06 19:56:49.459 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start  
2022-11-06 19:56:49.459 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir  
2022-11-06 19:56:49.459 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id  
2022-11-06 19:56:49.460 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id  
2022-11-06 19:56:49.460 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skippedrecords  
2022-11-06 19:56:49.460 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name  
2022-11-06 19:56:49.506 INFO streaming.PipeMapRed: R/W=S=1/B=0 in:NA [rec/s] out:NA [rec/s]  
2022-11-06 19:56:49.506 INFO streaming.PipeMapRed: R/W=S=1/B=0 in:NA [rec/s] out:NA [rec/s]  
2022-11-06 19:56:49.798 INFO streaming.PipeMapRed: Records R/W=72/3  
2022-11-06 19:56:49.798 INFO streaming.PipeMapRed: MapReduce step done  
2022-11-06 19:56:49.797 INFO streaming.PipeMapRed: mapredfinished  
2022-11-06 19:56:49.797 INFO mapred.LocalJobRunner:  
2022-11-06 19:56:49.797 INFO mapred.MapTask: Starting flush of map output  
2022-11-06 19:56:49.797 INFO mapred.MapTask: Spilling map output  
2022-11-06 19:56:49.797 INFO mapred.MapTask: bufstart = 0; bufend = 84; bufvoid = 104857600  
2022-11-06 19:56:49.797 INFO mapred.MapTask: kvstart = 26214396(104857584); length = 9/6553600
```

```

~/ECC_A1_code -- -zsh                                         /opt/homebrew/Cellar/hadoop/3.3.4/libexec/etc/hadoop -- ssh localhost +
Merged Map outputs=1
GC time elapsed (ms)=0
Total committed heap usage (bytes)=322961408
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Output Format Counters
  Bytes Written=84
2022-11-06 19:56:50,075 INFO mapred.LocalJobRunner: Finishing task: attempt_local1936192088_0001_r_000000_0
2022-11-06 19:56:50,075 INFO mapred.LocalJobRunner: reduce task executor complete.
2022-11-06 19:56:50,354 INFO mapreduce.Job: Job job_local1936192088_0001 running in uber mode : false
2022-11-06 19:56:50,354 INFO mapreduce.Job: map 10% reduce 10%
2022-11-06 19:56:50,354 INFO mapreduce.Job: Job job_local1936192088_0001 completed successfully
2022-11-06 19:56:50,368 INFO mapreduce.Job: Counters: 36
  File System Counters
    FILE: Number of bytes read=5988
    FILE: Number of bytes written=1315614
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=3784
    HDFS: Number of bytes written=84
    HDFS: Number of read operations=15
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=4
    HDFS: Number of bytes read erasure-coded=0
  Map-Reduce Framework
    Map input records=72
    Map output records=3
    Map output bytes=84
    Map output materialized bytes=96
    Input split bytes=115
    Combine input records=0
    Compressed input bytes=0
    Reduce input groups=3
    Reduce shuffle bytes=94
    Reduce input records=3
    Reduce output records=3
    Spilled Records=6
    Shuffled Maps =1
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=3
    Total committed heap usage (bytes)=645922816
  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=1892
  File Output Format Counters
    Bytes Written=84
2022-11-06 19:56:50,368 INFO streaming.StreamJob: Output directory: /user/ECC_A1-output_range100
(base) arpitawelling@Aripitas-MacBook-Pro ~ %

```

```

~/ECC_A1_code -- -zsh                                         /opt/homebrew/Cellar/hadoop/3.3.4/libexec/etc/hadoop -- ssh localhost +
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
  File Output Format Counters
    Bytes Written=84
2022-11-06 19:56:50,075 INFO mapred.LocalJobRunner: Finishing task: attempt_local1936192088_0001_r_000000_0
2022-11-06 19:56:50,075 INFO mapred.LocalJobRunner: reduce task executor complete.
2022-11-06 19:56:50,354 INFO mapreduce.Job: Job job_local1936192088_0001 running in uber mode : false
2022-11-06 19:56:50,354 INFO mapreduce.Job: map 10% reduce 10%
2022-11-06 19:56:50,354 INFO mapreduce.Job: Job job_local1936192088_0001 completed successfully
2022-11-06 19:56:50,368 INFO mapreduce.Job: Counters: 36
  File System Counters
    FILE: Number of bytes read=5988
    FILE: Number of bytes written=1315614
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=3784
    HDFS: Number of bytes written=84
    HDFS: Number of read operations=15
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=4
    HDFS: Number of bytes read erasure-coded=0
  Map-Reduce Framework
    Map input records=72
    Map output records=3
    Map output bytes=84
    Map output materialized bytes=96
    Input split bytes=115
    Combine output records=0
    Redistribute output bytes=0
    Reduce input groups=3
    Reduce shuffle bytes=94
    Reduce input records=3
    Reduce output records=3
    Spilled Records=6
    Shuffled Maps =1
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=3
    Total committed heap usage (bytes)=645922816
  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=1892
  File Output Format Counters
    Bytes Written=84
2022-11-06 19:56:50,368 INFO streaming.StreamJob: Output directory: /user/ECC_A1-output_range100
(base) arpitawelling@Aripitas-MacBook-Pro ~ % hdfs dfs -cat /user/ECC_A1-output_range100/part-00000
[10:00]151.239.241.163 9824
[09:00]151.239.241.163 9169
[08:00]151.239.241.163 6256
(base) arpitawelling@Aripitas-MacBook-Pro ~ %

```

Another question was to run 4 jobs together and use the capacity and fair scheduler. The 4 jobs that I ran were: Map reduce from Q1, Wordcount, Sort and Grep.

For Wordcount I gave my own input and did a map reduce to count how many times each word comes in the input. For Grep, I used the ip address regex and for sort, I sorted based on ip count.

The output of all these codes are given below:

MapReduce Word Count:

```
~/ECC_A1_code --zsh                                     /opt/homebrew/Cellar/hadoop/3.3.4/libexec/etc/hadoop - ssh localhost +  
[base] arpitawelling@Arpitas-MacBook-Pro ~ % /opt/homebrew/Cellar/hadoop/3.3.4/bin/hadoop jar /Users/arpitawelling/ECC_A1/hadoop-*streaming*.jar -file /Users/arpitawelling/ECC_A1_code/mapper_wordcount.py -mapper /Users/arpitawelling/ECC_A1_code/mapper_wordcount.py -file /Users/arpitawelling/ECC_A1_code/reducer_wordcount.py -reducer /Users/arpitawelling/ECC_A1_code/reducer_wordcount.py -input /user/wordcount1.txt -output /user/ECC_A1_output_wordcount10  
2022-11-20 08:01:41,379 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.  
2022-11-20 08:01:41,437 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-Java classes where applicable  
packageJobJar: /Users/arpitawelling/ECC_A1_code/mapper_wordcount.py, /Users/arpitawelling/ECC_A1_code/reducer_wordcount.py [] /var/folders/35/j_3438ps661dz9zwycq4_fdr000gn/T/streamjob3228029710811509924  
2022-11-20 08:01:41,744 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties  
2022-11-20 08:01:41,781 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).  
2022-11-20 08:01:41,781 INFO impl.MetricsSystemImpl: JobTracker metrics system started  
2022-11-20 08:01:41,787 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!  
2022-11-20 08:01:41,832 ERROR streaming.StreamJob: Error Launching job : Output directory hdfs://localhost:8020/user/ECC_A1_output_wordcount10 already exists  
Streaming Command Failed!  
[base] arpitawelling@Arpitas-MacBook-Pro ~ % /opt/homebrew/Cellar/hadoop/3.3.4/bin/hadoop jar /Users/arpitawelling/ECC_A1/hadoop-*streaming*.jar -file /Users/arpitawelling/ECC_A1_code/mapper_wordcount.py -mapper /Users/arpitawelling/ECC_A1_code/mapper_wordcount.py -file /Users/arpitawelling/ECC_A1_code/reducer_wordcount.py -reducer /Users/arpitawelling/ECC_A1_code/reducer_wordcount.py -input /user/wordcount1.txt -output /user/ECC_A1_output_wordcount10  
2022-11-20 08:01:47,957 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.  
2022-11-20 08:01:48,061 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-Java classes where applicable  
packageJobJar: /Users/arpitawelling/ECC_A1_code/mapper_wordcount.py, /Users/arpitawelling/ECC_A1_code/reducer_wordcount.py [] /var/folders/35/j_3438ps661dz9zwycq4_fdr000gn/T/streamjob98516653332845314  
2022-11-20 08:01:48,308 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties  
2022-11-20 08:01:48,345 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).  
2022-11-20 08:01:48,345 INFO impl.MetricsSystemImpl: JobTracker metrics system started  
2022-11-20 08:01:48,353 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!  
2022-11-20 08:01:48,476 INFO mapred.FileInputFormat: Total input files to process : 1  
2022-11-20 08:01:48,532 INFO mapreduce.JobSubmitter: number of splits:1  
2022-11-20 08:01:48,616 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1866231851_0001  
2022-11-20 08:01:48,616 INFO mapreduce.JobSubmitter: Executing with tokens: []  
2022-11-20 08:01:48,616 INFO mapreduce.Job: Job local1866231851_0001 running local file:/Users/arpitawelling/ECC_A1_code/mapper_wordcount.py as file:/opt/homebrew/Cellar/hadoop/hdfs/tmp/mapred/local/job_local1866231851_0001_jar_file_1708637459_4899_b38-1f658adfcf/mapper_wordcount.py  
2022-11-20 08:01:48,716 INFO mapred.LocalDistributedCacheManager: Local file:/Users/arpitawelling/ECC_A1_code/reducer_wordcount.py as file:/opt/homebrew/Cellar/hadoop/hdfs/tmp/mapred/local/job_local1866231851_0001_jar_file_4f3a-a12-65a347e42c75/reducer_wordcount.py  
2022-11-20 08:01:48,745 INFO mapreduce.Job: The url to track the job: http://localhost:8080/  
2022-11-20 08:01:48,746 INFO mapred.LocalJobRunner: OutputCommitter set in config null  
2022-11-20 08:01:48,746 INFO mapreduce.Job: Running job: job_local1866231851_0001  
2022-11-20 08:01:48,749 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter  
2022-11-20 08:01:48,749 INFO mapred.LocalJobRunner: OutputCommitter Algorithm version is 2  
2022-11-20 08:01:48,749 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false  
2022-11-20 08:01:48,804 INFO mapred.LocalJobRunner: Waiting for map tasks  
2022-11-20 08:01:48,816 INFO output.FileOutputCommitter: Starting task: attempt_local1866231851_0001_m_0000000_0  
2022-11-20 08:01:48,816 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2  
2022-11-20 08:01:48,819 INFO util.ProfsBasedProcessTree: ProfsBasedProcessTree currently is supported only on Linux.  
2022-11-20 08:01:48,819 INFO util.ProfsBasedProcessTree: Using org.apache.hadoop.mapred.util.PercentileCalculatorProcessTree : null  
2022-11-20 08:01:48,832 INFO mapred.Task: Map task 0 is starting  
2022-11-20 08:01:48,838 INFO mapred.Task: Map task 0 is assigned to hdfs://localhost:8020/user/wordcount1.txt:0+86  
2022-11-20 08:01:48,843 INFO mapred.MapTask: numReduceTasks: 1  
2022-11-20 08:01:48,843 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396/194857584  
2022-11-20 08:01:48,850 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100  
2022-11-20 08:01:48,843 INFO mapred.MapTask: soft limit at 83886080  
2022-11-20 08:01:48,843 INFO mapred.MapTask: bufstart = 0; bufwidth = 104857600  
2022-11-20 08:01:48,843 INFO mapred.MapTask: Kvstart = 26224396; length = 6553600  
2022-11-20 08:01:48,843 INFO mapred.MapTask: Map task 0 is assigned to hdfs://localhost:8020/user/wordcount1.txt  
2022-11-20 08:01:48,849 INFO mapred.MapTask: Map task 0 is assigned to hdfs://localhost:8020/user/wordcount1.py  
2022-11-20 08:01:48,851 INFO Configuration.deprecation: mapred.output.dir is deprecated. Instead, use mapreduce.task.output.dir  
2022-11-20 08:01:48,851 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir  
2022-11-20 08:01:48,852 INFO Configuration.deprecation: mapred.input.file is deprecated. Instead, use mapreduce.map.input.file  
2022-11-20 08:01:48,852 INFO Configuration.deprecation: mapred.input.length is deprecated. Instead, use mapreduce.map.input.length  
2022-11-20 08:01:48,852 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id  
2022-11-20 08:01:48,852 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition  
2022-11-20 08:01:48,853 INFO Configuration.deprecation: mapred.input.start is deprecated. Instead, use mapreduce.map.input.start  
2022-11-20 08:01:48,853 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
```

```

~/ECC_A1_code --zsh                                         /opt/homebrew/Cellar/hadoop/3.3.4/libexec/etc/hadoop -- ssh localhost +
Merged Map outputs=1
GC time elapsed (ms)=0
Total committed heap usage (bytes)=322961408
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Output Format Counters
  Bytes Written=46
2022-11-06 20:01:49,123 INFO mapred.LocalJobRunner: Finishing task: attempt_local866231851_0001_r_000000_0
2022-11-06 20:01:49,123 INFO mapred.LocalJobRunner: reduce task executor complete.
2022-11-06 20:01:49,754 INFO mapreduce.Job: Job job_local866231851_0001 running in uber mode : false
2022-11-06 20:01:49,754 INFO mapreduce.Job: map 10% reduce 100%
2022-11-06 20:01:49,766 INFO mapreduce.Job: Job job_local866231851_0001 completed successfully
2022-11-06 20:01:49,773 INFO mapreduce.Job: Counters: 36
  File System Counters
    FILE: Number of bytes read=3988
    FILE: Number of bytes written=1313732
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=172
    HDFS: Number of bytes written=46
    HDFS: Number of read operations=15
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=4
    HDFS: Number of bytes read erasure-coded=0
  Map-Reduce Framework
    Map input records=9
    Map output records=5
    Map output bytes=46
    Map output materialized bytes=62
    Input split bytes=93
    Combine input records=0
    Compressed input bytes=0
    Reduce input groups=5
    Reduce shuffle bytes=62
    Reduce input records=5
    Reduce output records=5
    Spilled Records=10
    Shuffled Maps =1
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=3
    Total committed heap usage (bytes)=645922816
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=86
  File Output Format Counters
  Bytes Written=46
2022-11-06 20:01:49,773 INFO streaming.StreamJob: Output directory: /user/ECC_A1-output_wordcount100
(base) arpitawelling@Arpitas-MacBook-Pro ~ %

```



```

~/ECC_A1_code --zsh                                         /opt/homebrew/Cellar/hadoop/3.3.4/libexec/etc/hadoop -- ssh localhost +
WRONG_REDUCE=0
File Output Format Counters
  Bytes Written=46
2022-11-06 20:01:49,123 INFO mapred.LocalJobRunner: Finishing task: attempt_local866231851_0001_r_000000_0
2022-11-06 20:01:49,123 INFO mapred.LocalJobRunner: reduce task executor complete.
2022-11-06 20:01:49,754 INFO mapreduce.Job: Job job_local866231851_0001 running in uber mode : false
2022-11-06 20:01:49,754 INFO mapreduce.Job: map 10% reduce 100%
2022-11-06 20:01:49,766 INFO mapreduce.Job: Job job_local866231851_0001 completed successfully
2022-11-06 20:01:49,773 INFO mapreduce.Job: Counters: 36
  File System Counters
    FILE: Number of bytes read=3988
    FILE: Number of bytes written=1313732
    FILE: Number of read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=172
    HDFS: Number of bytes written=46
    HDFS: Number of read operations=15
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=4
    HDFS: Number of bytes read erasure-coded=0
  Map-Reduce Framework
    Map input records=9
    Map output records=5
    Map output bytes=46
    Map output materialized bytes=62
    Input split bytes=93
    Combine input records=0
    Combine output records=0
    Reduce input groups=5
    Reduce shuffle bytes=62
    Reduce input records=5
    Reduce output records=5
    Spilled Records=10
    Shuffled Maps =1
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=3
    Total committed heap usage (bytes)=645922816
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=86
  File Output Format Counters
  Bytes Written=46
2022-11-06 20:01:49,773 INFO streaming.StreamJob: Output directory: /user/ECC_A1-output_wordcount100
(base) arpitawelling@Arpitas-MacBook-Pro ~ cat /user/ECC_A1-output_wordcount100/part-00000
cat: /user/ECC_A1-output_wordcount100/part-00000: No such file or directory
(base) arpitawelling@Arpitas-MacBook-Pro ~ hdfs dfs -cat /user/ECC_A1-output_wordcount100/part-00000
2022-11-06 20:02:42,946 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-Java classes where applicable
Aditi 2
Arpit 3
Harshitha 2
Sanika 2
Tani 3
(base) arpitawelling@Arpitas-MacBook-Pro ~ %

```

Map Reduce Sort:

```

~/ECC_A1_code -- -zsh                                         /opt/homebrew/Cellar/hadoop/3.3.4/libexec/etc/hadoop -- ssh localhost +
```

```

(base) arpitawelling@Aritpatas-MacBook-Pro: ~ % /opt/homebrew/Cellar/hadoop/3.3.4/bin/hadoop jar /Users/arpitawelling/ECC_A1_code/mapper_sort.py -file /Users/arpitawelling/ECC_A1_code/reducer_sort.py -reducer /Users/arpitawelling/ECC_A1_code/reducer_sort.py -input /user/ip_data.txt -output /user/ECC_A1-output_sort100
2022-11-06 20:04:06,315 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
2022-11-06 20:04:06,374 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-Java classes where applicable
packageJobJar: [/Users/arpitawelling/ECC_A1_code/mapper_sort.py, /Users/arpitawelling/ECC_A1_code/reducer_sort.py] [] /var/folders/35/j_3430ps66ldzrwycq4.fdr:0000gn/T/streamjob3015330868148185214.jar tmp0
ir=multi
2022-11-06 20:04:06,499 INFO Impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2022-11-06 20:04:06,728 INFO Impl.MetricsSystemImpl: Scheduled Metrics snapshot period = 10 seconds.
2022-11-06 20:04:06,728 INFO Impl.MetricsSystemImpl: JobTracker metrics system started
2022-11-06 20:04:06,862 INFO mapred.FileInputFormat: Total input files to process : 1
2022-11-06 20:04:06,922 INFO mapreduce.JobSubmitter: number of splits:1
2022-11-06 20:04:07,004 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local285941331_0001
2022-11-06 20:04:07,084 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-11-06 20:04:07,144 INFO mapreduce.JobSubmitter: Localized file:/Users/arpitawelling/ECC_A1_code/mapper_sort.py as file:/opt/homebrew/Cellar/hadoop/tmp/mapred/local/job_local285941331_0001_c091966-71aa-bd9d-1c74c3bb94dd/mapper_sort.py
2022-11-06 20:04:07,154 INFO mapred.LocalDistributedCacheManager: Localized file:/Users/arpitawelling/ECC_A1_code/reducer_sort.py as file:/opt/homebrew/Cellar/hadoop/tmp/mapred/local/job_local285941331_0001_37d9fb0b-b9b3-475b-86c9-c9cd77789bc/reducer_sort.py
2022-11-06 20:04:07,164 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2022-11-06 20:04:07,147 INFO mapreduce.Job: Running job: job_local285941331_0001
2022-11-06 20:04:07,148 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2022-11-06 20:04:07,148 INFO mapred.LocalJobRunner: org.apache.hadoop.mapred.FileOutputCommitter
2022-11-06 20:04:07,158 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2022-11-06 20:04:07,158 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2022-11-06 20:04:07,207 INFO mapred.LocalJobRunner: Waiting for map tasks
2022-11-06 20:04:07,208 INFO mapred.LocalJobRunner: Starting task: attempt_local285941331_0001_m_000000.0
2022-11-06 20:04:07,218 INFO mapred.LocalJobRunner: Starting task: attempt_local285941331_0001_m_000000.0
2022-11-06 20:04:07,222 INFO mapred.FileOutputCommitter: File Output Committer skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2022-11-06 20:04:07,222 INFO mapred.FileOutputCommitter: File Output Committer currently is supported only on Linux.
2022-11-06 20:04:07,222 INFO mapred.Task: Using ResourceCalculatorProcessTree : null
2022-11-06 20:04:07,224 INFO mapred.Task: Processing split:hdfs://localhost:8020/user/ip_data.txt:0+21499
2022-11-06 20:04:07,234 INFO mapred.MapTask: numReduceTasks: 1
2022-11-06 20:04:07,254 INFO mapred.MapTask: (EQUATOR) 0 kv: 26214396((104857584)
2022-11-06 20:04:07,256 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 108
2022-11-06 20:04:07,256 INFO mapred.MapTask: soft limit at 8386898
2022-11-06 20:04:07,256 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2022-11-06 20:04:07,256 INFO mapred.MapTask: length = 655360
2022-11-06 20:04:07,258 INFO mapred.MapTask: map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2022-11-06 20:04:07,264 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/arpitawelling/.mapper_sort.py]
2022-11-06 20:04:07,267 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
2022-11-06 20:04:07,267 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
2022-11-06 20:04:07,267 INFO Configuration.deprecation: mapr.input.file is deprecated. Instead, use mapreduce.map.input.file
2022-11-06 20:04:07,267 INFO Configuration.deprecation: mapr.input.length is deprecated. Instead, use mapreduce.map.input.length
2022-11-06 20:04:07,267 INFO Configuration.deprecation: mapr.input.start is deprecated. Instead, use mapreduce.map.input.start
2022-11-06 20:04:07,268 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
2022-11-06 20:04:07,268 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
2022-11-06 20:04:07,269 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
2022-11-06 20:04:07,269 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
2022-11-06 20:04:07,269 INFO Configuration.deprecation: mapred.skip.threshold is deprecated. Instead, use mapreduce.task.partition
2022-11-06 20:04:07,270 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
2022-11-06 20:04:07,270 INFO Configuration.deprecation: mapred.input.start is deprecated. Instead, use mapreduce.map.input.start
2022-11-06 20:04:07,270 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
2022-11-06 20:04:07,270 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
2022-11-06 20:04:07,270 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
2022-11-06 20:04:07,321 INFO streaming.PipeMapRed: R/W=S+1/V/B inNA [rec/s] outNA [rec/s]
2022-11-06 20:04:07,321 INFO streaming.PipeMapRed: R/W=S+10/0/0 inNA [rec/s] outNA [rec/s]
2022-11-06 20:04:07,323 INFO streaming.PipeMapRed: Records R/W=72/7
2022-11-06 20:04:07,324 INFO streaming.PipeMapRed: MRErrorThread done
2022-11-06 20:04:07,324 INFO streaming.PipeMapRed: mapRedefined
2022-11-06 20:04:07,325 INFO mapred.LocalJobRunner: 
2022-11-06 20:04:07,325 INFO mapred.MapTask: Starting flush of map output
2022-11-06 20:04:07,325 INFO mapred.MapTask: Map output
2022-11-06 20:04:07,325 INFO mapred.MapTask: burstart = 0; bufend = 930; bufvoid = 104857600

```

```

~/ECC_A1_code -- -zsh                                         /opt/homebrew/Cellar/hadoop/3.3.4/libexec/etc/hadoop -- ssh localhost +
```

```

Merged Map outputs=1
GC time elapsed (ms)=0
Total committed heap usage (bytes)=322961408
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Written=281
2022-11-06 20:04:07,549 INFO mapred.LocalJobRunner: Finishing task: attempt_local285941331_0001_r_000000_0
2022-11-06 20:04:07,549 INFO mapred.LocalJobRunner: reduce task executor complete.
2022-11-06 20:04:08,158 INFO mapreduce.Job: Job: job_local285941331_0001 running in uber mode : false
2022-11-06 20:04:08,158 INFO mapreduce.Job: map 100% reduce 100%
2022-11-06 20:04:08,158 INFO mapreduce.Job: Job: job_local285941331_0001 completed successfully
2022-11-06 20:04:08,158 INFO mapreduce.Job: Counters
  File System Counters
    FILE: Number of bytes read=6184
    FILE: Number of bytes written=1316850
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=2299
    HDFS: Number of bytes written=281
    HDFS: Number of read operations=15
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=4
    HDFS: Number of bytes read erasure-coded=0
  Map-Reduce Framework
    Map input records=72
    Map output records=72
    Map output bytes=938
    Map output materialized bytes=1088
    Input split bytes=98
    Combine input records=0
    Combine output records=0
    Reduce input groups=1
    Reduce input bytes=1088
    Reduce input records=72
    Reduce output records=22
    Spilled Records=144
    Shuffled Maps =1
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=4
    Total committed heap usage (bytes)=645922816
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=21499
  File Output Format Counters
  Bytes Written=281
2022-11-06 20:04:08,165 INFO streaming.StreamJob: Output directory: /user/ECC_A1-output_sort100
(base) arpitawelling@Aritpatas-MacBook-Pro: ~ %

```

```

~/ECC_A1_code -- -zsh
HDFS: Number of bytes read=9988
HDFS: Number of bytes written=281
HDFS: Number of read operations=15
HDFS: Number of large read operations=0
HDFS: Number of write operations=4
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=72
  Map output records=72
  Map output bytes=938
  Map output materialized bytes=1080
  Input split bytes=98
  Combine input records=0
  Combine output records=0
  Reduce input groups=22
  Reduce input bytes=9880
  Reduce input records=72
  Reduce output records=22
  Spilled Records=144
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  Data relayout elapsed (ms)=4
  Total committed heap usage (bytes)=645922816
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_SOURCE=0
File Input Format Counters
  Bytes Read=21499
File Output Format Counters
  Bytes Written=281
2022-11-06 20:04:08,165 INFO streaming.StreamJob: Output directory: /user/ECC_A1-output_sort100
(base) arpitawelling@Arpitas-MacBook-Pro ~ % hdfs dfs -cat /user/ECC_A1-output_sort100/part-00000
2022-11-06 20:04:08,556 WARN Util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
127.220.71 1
13.66.139 1
151.242.186 1
192.15.51 1
210.156.101 1
238.111.16 1
43.157.17 1
44.154.188 1
54.36.149 1
91.185.153 1
91.99.55 1
178.253.33 2
287.46.13 2
66.149.49 3
91.99.72 3
187.155.181 4
31.56.96 4
120.204.67 5
210.158.93 5
117.177.254 10
48.77.167 10
185.158.236 13
(base) arpitawelling@Arpitas-MacBook-Pro ~ %

```

Map Reduce grep:

```
-@ECC_A1_code -- -zsh
(base) arpitawelling@Arpitas-MacBook-Pro ~ % /opt/homebrew/Cellar/hadoop/3.3.4/bin/hadoop jar /Users/arpitawelling/ECC_A1_code/mapper_grep.py -file /Users/arpitawelling/ECC_A1_code/mapper_grep.py -reducer /Users/arpitawelling/ECC_A1_code/reducer_grep.py -input /user/grep_input.txt -output /user/ECC_A1-output_grep100
2022-11-06 20:05:49,051 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
2022-11-06 20:05:49,015 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-Java classes where applicable
packageJobJar: [/Users/arpitawelling/ECC_A1_code/mapper_grep.py, /Users/arpitawelling/ECC_A1_code/reducer_grep.py] [] /var/folders/35/j_3430ps66ldzrzwycq4.fdr:0000gn/T/streamjob11763199152827847197.jar tmp
Dir=null
2022-11-06 20:05:50,236 INFO Impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2022-11-06 20:05:50,272 INFO Impl.MetricsSystemImpl: Scheduled Metric snapshot period = 10 seconds.
2022-11-06 20:05:50,274 INFO Impl.MetricsSystemImpl: JobTracker metrics system started
2022-11-06 20:05:50,407 INFO mapred.FileInputFormat: Total input files to process : 1
2022-11-06 20:05:50,469 INFO mapreduce.JobSubmitter: number of splits:1
2022-11-06 20:05:50,548 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1311494483_0001
2022-11-06 20:05:50,548 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-11-06 20:05:50,548 INFO mapreduce.JobSubmitter: LocalDistributedCacheManager: Localized file:/Users/arpitawelling/ECC_A1_code/mapper_grep.py as file:/opt/homebrew/Cellar/hadoop/tmp/mapred/local/job_local1311494483_0001_1185471c-fc8-4f2c-95b1-2335ff110224/mapper_grep.py
2022-11-06 20:05:50,548 INFO mapred.LocalDistributedCacheManager: Localized file:/Users/arpitawelling/ECC_A1_code/reducer_grep.py as file:/opt/homebrew/Cellar/hadoop/tmp/mapred/local/job_local1311494483_0001_118541c5-f29b-477e-a378-9b1c122b5273/reducer_grep.py
2022-11-06 20:05:50,685 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2022-11-06 20:05:50,684 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2022-11-06 20:05:50,684 INFO mapreduce.Job: Running Job: job_local1311494483_0001
2022-11-06 20:05:50,684 INFO mapred.LocalJobRunner: OutputCommitter org.apache.hadoop.mapred.FileOutputCommitter
2022-11-06 20:05:50,689 INFO mapred.FileOutputCommitter: File Output Committer Algorithm version is 2
2022-11-06 20:05:50,689 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2022-11-06 20:05:50,728 INFO mapred.LocalJobRunner: Waiting for map tasks
2022-11-06 20:05:50,728 INFO mapred.LocalJobRunner: Starting task: attempt_local1311494483_0001_m_000000_0
2022-11-06 20:05:50,739 INFO mapred.FileOutputCommitter: File Output Committer Algorithm version is 2
2022-11-06 20:05:50,739 INFO mapred.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2022-11-06 20:05:50,742 INFO mapred.Task: Using ResourceCalculatorProcessTree : currently is supported only on Linux.
2022-11-06 20:05:50,742 INFO mapred.Task: Processing split:hdfs://localhost:8020/user/grep_input.txt:0+1242
2022-11-06 20:05:50,753 INFO mapred.MapTask: numReduceTasks: 1
2022-11-06 20:05:50,774 INFO mapred.MapTask: (EQUATOR) 0 kv: 26214396((104857584)
2022-11-06 20:05:50,774 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 108
2022-11-06 20:05:50,776 INFO mapred.MapTask: soft limit at 83868698
2022-11-06 20:05:50,776 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2022-11-06 20:05:50,776 INFO mapred.MapTask: length = 6553600
2022-11-06 20:05:50,777 INFO mapred.MapTask: map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2022-11-06 20:05:50,782 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/arpitawelling/.mapper_grep.py]
2022-11-06 20:05:50,783 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
2022-11-06 20:05:50,784 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
2022-11-06 20:05:50,784 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
2022-11-06 20:05:50,784 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
2022-11-06 20:05:50,784 INFO Configuration.deprecation: map.input.location is deprecated. Instead, use mapreduce.map.input.location
2022-11-06 20:05:50,784 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
2022-11-06 20:05:50,784 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
2022-11-06 20:05:50,785 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
2022-11-06 20:05:50,785 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
2022-11-06 20:05:50,785 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
2022-11-06 20:05:50,787 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
2022-11-06 20:05:50,788 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.user.name
2022-11-06 20:05:50,834 INFO streaming.PipeMapRed: R/W=S+I/V=0 inNA [rec/s] outNA [rec/s]
2022-11-06 20:05:50,834 INFO streaming.PipeMapRed: R/W=S+I/O/O inNA [rec/s] outNA [rec/s]
2022-11-06 20:05:50,834 INFO streaming.PipeMapRed: Records R/W=10/
2022-11-06 20:05:50,834 INFO streaming.PipeMapRed: MRErrorThread done
2022-11-06 20:05:50,834 INFO streaming.PipeMapRed: mapRedefined
2022-11-06 20:05:50,837 INFO mapred.LocalJobRunner:
2022-11-06 20:05:50,837 INFO mapred.MapTask: Starting flush of map output
2022-11-06 20:05:50,837 INFO mapred.MapTask: Spilling map output
2022-11-06 20:05:50,837 INFO mapred.MapTask: burstart = 0; bufend = 1168; bufvoid = 104857600
```

```
-@ECC_A1_code -- -zsh
HDFS: Number of bytes written=1181
HDFS: Number of read operations=15
HDFS: Number of large read operations=0
HDFS: Number of write operations=4
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
  Map input records=10
  Map output records=10
  Map output bytes=1168
  Map output compressed bytes=1201
  Input split bytes=93
  Combine input records=0
  Combine output records=0
  Reduce input groups=6
  Reduce shuffle bytes=1201
  Reduce input records=10
  Reduce output records=20
  Spilled Maps=20
  Shuffled Maps=1
  Failed Shuffiles=0
  Merged Map outputs=1
  GC time elapsed (ms)=3
  Total committed heap usage (bytes)=645922816
Shuffle Errors
  Invalid ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=112
  File Output Format Counters
  Bytes Written=1181
2022-11-06 20:05:51,711 INFO streaming.StreamJob: Output directory: /user/ECC_A1-output_grep100
(base) arpitawelling@Arpitas-MacBook-Pro ~ % hdfs dfs -cat /user/ECC_A1-output_grep100/part-00000
2022-11-06 20:07:16,227 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-Java classes where applicable
```

```
31.56.96.51 -- [22/Jan/2019:03:56:16 +0330] "GET /image/68844/productModel/200x200 HTTP/1.1" 200 5667 "https://www.zanbil.ir/m/filter/bl13" "Mozilla/5.0 (Linux; Android 6.0; ALE-L21 Build/HuaweiALE-L21) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/66.0.3359.158 Mobile Safari/537.36" "-"
31.56.96.51 -- [22/Jan/2019:03:56:16 +0330] "GET /image/01474/productModel/200x200 HTTP/1.1" 200 5379 "https://www.zanbil.ir/m/filter/bl13" "Mozilla/5.0 (Linux; Android 6.0; ALE-L21 Build/HuaweiALE-L21) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/66.0.3359.158 Mobile Safari/537.36" "-"
48.77.167.129 -- [22/Jan/2019:03:56:17 +0330] "GET /image/14925/productModel/100x100 HTTP/1.1" 200 1696 "-" "Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)" "-"
54.36.149.41 -- [22/Jan/2019:03:56:14 +0330] "GET /filter/2713b28d95b85d0d8a7nd98remdbn85dn49d8b3nd9684,2710d4av9nd95b85dn49db812nd8847nd8nb285%20d9b5nd49af0d8b7nd98en8db8cnd8a9nd8nb3d984,p83 HTTP/1.1" 200 30577 "-" "Mozilla/5.0 (compatible; AhrefsBot/6.1; +http://ahrefs.com/robot/)" "-"
91.99.72.16 -- [22/Jan/2019:03:56:17 +0330] "GET /product/
(base) arpitawelling@Arpitas-MacBook-Pro ~ %
```

Capacity Scheduler:

After running 4 parallel jobs and using Capacity Scheduler, it is seen that a capacity scheduler takes one job and other jobs have to wait till the current job is running.

The outputs are seen below.

Screenshots for Capacity scheduler:

n	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Allocated GPUs	Reserved CPU VCores	Reserved Memory MB	Reserved GPUs	% of Queue	% of Cluster	Progress	Tracking UI
	Sun Nov 6 18:29:23 -0500 2022	N/A	N/A	ACCEPTED	UNDEFINED	0	0	0	-1	0	0	-1	0.0	0.0	<div style="width: 0%;"></div>	ApplicationMaster
	Sun Nov 6 18:29:20 -0500 2022	N/A	N/A	ACCEPTED	UNDEFINED	0	0	0	-1	0	0	-1	0.0	0.0	<div style="width: 0%;"></div>	ApplicationMaster
	Sun Nov 6 18:29:14 -0500 2022	N/A	N/A	ACCEPTED	UNDEFINED	0	0	0	-1	0	0	-1	0.0	0.0	<div style="width: 0%;"></div>	ApplicationMaster
	Sun Nov 6 18:29:07 -0500 2022	Sun Nov 6 18:29:07 -0500 2022	N/A	RUNNING	UNDEFINED	4	4	5120	-1	0	0	-1	62.5	62.5	<div style="width: 62.5%; background-color: #cccccc;"></div>	ApplicationMaster

Dump scheduler logs 1 min ▾															
Application Queues															
Legend: Capacity Used Used (over capacity) Max Capacity Users Requesting Resources Auto Created Queues															
															
Show 20 ▾ entries															
ID	User	Name		Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Alloc CPU VCores	Alloc GPU VCores
application_1667745597871_0037	hadoop	streamjob8258080702587505468.jar		MAPREDUCE		default	0	Sun Nov 6 18:31:52 -0500 2022	N/A	N/A	ACCEPTED	UNDEFINED	0	0	0
application_1667745597871_0036	hadoop	streamjob1549685868557784652.jar		MAPREDUCE		default	0	Sun Nov 6 18:29:23 -0500 2022	N/A	N/A	ACCEPTED	UNDEFINED	0	0	0
application_1667745597871_0035	hadoop	streamjob8944103970597097878.jar		MAPREDUCE		default	0	Sun Nov 6 18:29:20 -0500 2022	N/A	N/A	ACCEPTED	UNDEFINED	0	0	0

Dump scheduler logs 1 min ▾															
Application Queues															
Legend: Capacity Used Used (over capacity) Max Capacity Users Requesting Resources Auto Created Queues															
															
Show 20 ▾ entries															
ID	User	Name		Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Alloc CPU VCores	Alloc GPU VCores
application_1667745597871_0037	hadoop	streamjob8258080702587505468.jar		MAPREDUCE		default	0	Sun Nov 6 18:31:52 -0500 2022	N/A	N/A	ACCEPTED	UNDEFINED	0	0	0
application_1667745597871_0036	hadoop	streamjob1549685868557784652.jar		MAPREDUCE		default	0	Sun Nov 6 18:29:23 -0500 2022	N/A	N/A	ACCEPTED	UNDEFINED	0	0	0

Auto Created Queues

25.0% used
25.0% used

Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Allocated GPUs	Reserved CPU VCores	Reserved Memory MB	Reserved GPUs	% of Queue	% of Cluster	Progress	Tracking UI
0	Sun Nov 6 18:31:52 -0500 2022	N/A	N/A	ACCEPTED	UNDEFINED	0	0	0	-1	0	0	-1	0.0	0.0		Application
0	Sun Nov 6 18:29:23 -0500 2022	N/A	N/A	ACCEPTED	UNDEFINED	0	0	0	-1	0	0	-1	0.0	0.0		Application

Dump scheduler logs | 1 min ▾

Application Queues

Legend: Used Used (over capacity) Max Capacity Users Requesting Resources Auto Created Queues

- Cluster root
- Queue detail

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores
application_166774559781_0037	hadoop	streamjob8258080702587505468.jar	MAPREDUCE		default	0	Sun Nov 6 18:31:52 -0500 2022	N/A	N/A	ACCEPTED	UNDEFINED	0	0

Showing 1 to 1 of 1 entries

Fair Scheduler:

For a fair scheduler, it can be seen that it allocates resources to one job and then after some progress, it will allocate the resources to another job and get back to the first job later. This helps in running the jobs in parallel as other jobs do not have to wait for earlier jobs to finish.

Screenshots for Fair scheduler:

Used Used (over fair share) Max Capacity

50.0% used
0.0% used
50.0% used

	Application Type	Queue	Fair Share	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	Progress	Tracking UI
!07824809.jar	MAPREDUCE	root.hadoop	2731	1667779633884	N/A	N/A	undefined undefined NaN NaN:NaN:NaN -0NaN NaN	UNDEFINED	0	0	0	0	0		UNASSIGNED
!153161182.jar	MAPREDUCE	root.hadoop	2731	1667779627567	Sun Nov 6 19:07:08 -0500 2022	N/A	undefined undefined NaN NaN:NaN:NaN -0NaN NaN	UNDEFINED	1	1	2048	0	0		UNASSIGNED
!307631907.jar	MAPREDUCE	root.hadoop	2731	1667779622540	Sun Nov 6 19:07:03 -0500 2022	N/A	undefined undefined NaN NaN:NaN:NaN -0NaN NaN	UNDEFINED	1	1	2048	0	0		UNASSIGNED

First Previous 1 Next Last

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster										
Fair Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0										
Application Queues														
Legend: Steady Fair Share Instantaneous Fair Share Used Used (over fair share) Max Capacity														
<ul style="list-style-type: none"> root + root.default + root.hadoop 														
Show 20 entries														
ID	User	Name	Application Type	Queue										
application_1667779415638_0004	hadoop	streamjob490046636207824809.jar	MAPREDUCE	root.hadoop										
application_1667779415638_0003	hadoop	streamjob8217907045153161182.jar	MAPREDUCE	root.hadoop										
application_1667779415638_0002	hadoop	streamjob1658958132307631907.jar	MAPREDUCE	root.hadoop										

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster										
Fair Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0										
Application Queues														
Legend: Steady Fair Share Instantaneous Fair Share Used Used (over fair share) Max Capacity														
<ul style="list-style-type: none"> root + root.default + root.hadoop 														
Show 20 entries														
ID	User	Name	Application Type	Queue										
application_1667779415638_0006	hadoop	streamjob7167982802033589443.jar	MAPREDUCE	root.hadoop										
application_1667779415638_0005	hadoop	streamjob525453148426938559.jar	MAPREDUCE	root.hadoop										

Search:														
Application Type	Queue	Fair Share	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU Vcores	Reserved Memory MB	Progress	Tracking URL
MAPREDUCE	root.hadoop	4096	1667779807034	Sun Nov 6 19:10:07 -0500 2022	N/A	undefined undefined NaN NaN:NaN:NaN -0NaN NaN	UNDEFINED	3	3	4096	0	0	<div style="width: 100%;"></div>	ApplicationMaster
MAPREDUCE	root.hadoop	4096	1667779638245	Sun Nov 6 19:10:02 -0500 2022	N/A	undefined undefined NaN NaN:NaN:NaN -0NaN NaN	UNDEFINED	2	2	3072	0	0	<div style="width: 50%;"></div>	ApplicationMaster
First Previous <input type="button" value="1"/> Next Last														
Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster										
Fair Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0										
Application Queues														
Legend: Steady Fair Share Instantaneous Fair Share Used Used (over fair share) Max Capacity														
<ul style="list-style-type: none"> root + root.default + root.hadoop 														
Show 20 entries														
ID	User	Name	Application Type	Queue										
application_1667779415638_0005	hadoop	streamjob525453148426938559.jar	MAPREDUCE	root.hadoop										

Challenges and testing environment:

I tried to do this assignment on Jetstream2 but for some reason, when I ran a job it used to get stuck and some point. I tried to debug this but was not successful. I eventually tested on my

laptop (MAC OS).

I discussed with my peers (Sanika Anil Paranjpe (sparanjp), Aditi Vijay Gode (adigode), Nithin Nataraj Vusirikala (nnvusiri)) on a high level about this assignment.

References:

<https://towardsdatascience.com/installing-hadoop-on-a-mac-ec01c67b003c>

<https://prwatech.in/blog/hadoop/hadoop-fair-scheduler-tutorial/>

<https://intellipaat.com/blog/tutorial/hadoop-tutorial/hdfs-operations/#:~:text=Datanode%20daemon%20should%20be%20started,identified%20by%20script%2Dbased%20commands.>

<https://medium.com/beeranddiapers/installing-hadoop-on-mac-a9a3649dbc4d>

<https://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/>