

E516 - ECC Assignment 1 (Fall '22)

Submitted by : Aniket Kale (ankale@iu.edu)

Setup

1. Create Jetstream 2 instance (medium-allocation). Connect using SSH to Jetstream 2.
2. Make user "hadoopuser".
3. Give sudo access to "hadoop user".
4. SSH to local-host.
5. Download Hadoop latest version and configure using [this](#) tutorial.
6. Start using command "start-dfs.sh" and "start-yarn.sh"
7. Run command "jps" to check if all service are running

```
[hadoopuser@assignment1-ankale-main:~$ jps
17091 SecondaryNameNode
29364 Jps
17415 ResourceManager
16648 NameNode
17582 NodeManager
16830 DataNode
hadoopuser@assignment1-ankale-main:~$ ]
```

Test wordcount program:

```
[root@node1 ~]# hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar -D mapreduce.job.reduces=1 wordcount /word_count_in_python/part-00000 /word_count_in_python/output
2022-11-03 01:13:03,334 INFO streaming.StreamJob: Output directory: /word_count_in_python/output
[hadoopuser@assignment1-ankale-main:~$ hdfs dfs -cat /word_count_in_python/output/part-00000
GeeksforGeeks 1
Hello 2
I 2
Intern 1
am 2
an 1
hadoopuser@assignment1-ankale-main:~$ ]
```

Part 1:

Get top 3 IPs for every hour from the log files.

1.1 Run map-reduce for top 3 on "sample.log":

- Move "sample.log", "mapper-ip.py", "reducer-main.py" to Jetstream2
- Copy "sample.log" to HDFS:

```
|hadoopuser@assignment1-ankale-main:~/part1$ ls
mapper-ip.py reducer-main.py sample.log
|hadoopuser@assignment1-ankale-main:~/part1$ cd ..
|hadoopuser@assignment1-ankale-main:~$ cd part1/
|hadoopuser@assignment1-ankale-main:~/part1$ chmod 777 mapper-ip.py reducer-main.py sample.log
|hadoopuser@assignment1-ankale-main:~/part1$ cd ..
|hadoopuser@assignment1-ankale-main:~$ hdfs dfs -mkdir smallLog
mkdir: `hdfs://127.0.0.1:9000/user/hadoopuser': No such file or directory
|hadoopuser@assignment1-ankale-main:~$ hdfs dfs -mkdir /smallLog
|hadoopuser@assignment1-ankale-main:~$ hdfs dfs -ls
ls: `.': No such file or directory
|hadoopuser@assignment1-ankale-main:~$ hdfs dfs -ls /smallLog
|hadoopuser@assignment1-ankale-main:~$ hdfs dfs -copyFromLocal /home/hadoopuser/part1/sample.log /smallLog
|hadoopuser@assignment1-ankale-main:~$ hdfs dfs -ls /smallLog
Found 1 items
-rw-r--r--    1 hadoopuser supergroup      102399 2022-11-05 17:17 /smallLog/sample.log
|hadoopuser@assignment1-ankale-main:~$ |
```

- Run the following command:

```
> hadoop jar
/home/hadoopuser/hadoop-3.3.4/share/hadoop/tools/lib/hadoop-streaming-3.3.
4.jar
-input /smallLog/sample.log
-output /smallLog/output -mapper "python3
/home/hadoopuser/part1/mapper-ip.py"
-reducer "python3 /home/hadoopuser/part1/reducer-main.py"
```

d. Run command:

```
hadoopuser@assignment1-ankale-main:~$ hadoop jar /home/hadoopuser/hadoop-3.3.4/share/hadoop/tools/lib/hadoop-streaming-3.3.4.jar  
-input /smallLog/sample.log -output /smallLog/output -mapper "python3 /home/hadoopuser/part1/mapper-ip.py" -reducer "python3  
/home/hadoopuser/part1/reducer-main.py"  
packageJobJar: [/tmp/hadoop-unjar7987586106297474119/] [] /tmp/streamjob2315091795510642602.jar tmpDir=null  
2022-11-05 17:29:30,371 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /127.0.0.1:8032  
2022-11-05 17:29:30,510 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /127.0.0.1:8032  
2022-11-05 17:29:30,698 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoopus  
er/.staging/job_1667431627519_0009  
2022-11-05 17:29:30,894 INFO mapred.FileInputFormat: Total input files to process : 1  
2022-11-05 17:29:31,343 INFO mapreduce.JobSubmitter: number of splits:2  
2022-11-05 17:29:31,584 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1667431627519_0009  
2022-11-05 17:29:31,642 INFO conf.Configuration: resource-types.xml not found  
2022-11-05 17:29:31,643 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.  
2022-11-05 17:29:31,706 INFO impl.YarnImpl: Submitted application application_1667431627519_0009  
2022-11-05 17:29:31,737 INFO mapreduce.Job: The url to track the job: http://assignment1-ankale-main.js2local:8088/proxy/applica  
tion_1667431627519_0009/  
2022-11-05 17:29:31,739 INFO mapreduce.Job: Running job: job_1667431627519_0009  
2022-11-05 17:29:36,794 INFO mapreduce.Job: Job job_1667431627519_0009 running in uber mode : false  
2022-11-05 17:29:36,795 INFO mapreduce.Job: map 0% reduce 0%
```

e. Show output:

```
Peak Map Virtual memory (bytes)=273332864  
Peak Reduce Physical memory (bytes)=226676736  
Peak Reduce Virtual memory (bytes)=2743332864  
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=106495  
File Output Format Counters  
Bytes Written=58  
2022-11-05 17:29:45,928 INFO streaming.StreamJob: Output directory: /smallLog/output  
hadoopuser@assignment1-ankale-main:~$ hdfs dfs -ls /smallLog/output  
Found 2 items  
-rw-r--r-- 1 hadoopuser supergroup 0 2022-11-05 17:29 /smallLog/output/_SUCCESS  
-rw-r--r-- 1 hadoopuser supergroup 58 2022-11-05 17:29 /smallLog/output/part-00000  
hadoopuser@assignment1-ankale-main:~$ hdfs dfs -cat /smallLog/output/part-00000  
03 66.111.54.249 38  
03 5.211.97.39 36  
03 66.249.66.194 31  
hadoopuser@assignment1-ankale-main:~$
```

1.2 Run map-reduce for top 3 on "access.log" (3.5 GB file):

a. Move "access.log" to Jetstream2 and copy to HDFS

```
hadoopuser@assignment1-ankale-main:~$ hdfs dfs -copyFromLocal /home/hadoopuser/part1/access.log /accessLog/access.log  
hadoopuser@assignment1-ankale-main:~$ hdfs dfs -ls /accessLog/  
Found 1 items  
-rw-r--r-- 1 hadoopuser supergroup 3502440823 2022-11-06 15:56 /accessLog/access.log  
hadoopuser@assignment1-ankale-main:~$
```

b. Run MapReduce

```

hadoopuser@assignment1-ankale-main:~$ hadoop jar /home/hadoopuser/hadoop-3.3.4/share/hadoop/tools/lib/hadoop-streaming-3.3.4.jar -input /accessLog/access.log -output /accessLog/output -mapper "python /home/hadoopuser/part1/mapper-ip.py" -reducer "python3 /home/hadoopuser/part1/reducer-main.py"
packageJobJar: [/tmp/hadoop-unjar1133877149428562375/] []
/tmp/streamjob2895926519263414969.jar tmpDir=null
2022-11-06 16:01:34,945 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /127.0.0.1:8082
2022-11-06 16:01:34,968 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /127.0.0.1:8082
2022-11-06 16:01:34,983 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoopuser/.staging/job_1667431627519_0010
0010
2022-11-06 16:01:34,985 INFO mapred.FileInputFormat: Total input files to process : 1
2022-11-06 16:01:34,987 INFO net.NetworkTopology: Adding a new node: /default-rack/127.0.0.1:9866
2022-11-06 16:01:34,987 INFO mapreduce.JobSubmitter: number of splits:26
2022-11-06 16:01:34,988 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1667431627519_0010
2022-11-06 16:01:34,988 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-11-06 16:01:34,989 INFO conf.Configuration: resource-types.xml not found
2022-11-06 16:01:34,989 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-11-06 16:01:34,989 INFO impl.YarnClientImpl: Submitted application application_1667431627519_0010
2022-11-06 16:01:34,989 INFO mapreduce.Job: The url to track the job: http://assignment1-ankale-main.js2local:8088/proxy/application_1667431627519_0010/
2022-11-06 16:01:34,989 INFO mapreduce.Job: Running job: job_1667431627519_0010 running in uber mode : false
2022-11-06 16:01:34,989 INFO mapreduce.Job: map 0% reduce 0%
2022-11-06 16:01:49,926 INFO mapreduce.Job: map 8% reduce 0%
2022-11-06 16:01:57,971 INFO mapreduce.Job: map 14% reduce 0%
2022-11-06 16:02:03,998 INFO mapreduce.Job: map 17% reduce 0%
2022-11-06 16:02:05,905 INFO mapreduce.Job: map 19% reduce 0%
2022-11-06 16:02:06,113 INFO mapreduce.Job: map 21% reduce 0%
2022-11-06 16:02:11,046 INFO mapreduce.Job: map 22% reduce 0%
2022-11-06 16:02:12,058 INFO mapreduce.Job: map 26% reduce 0%
2022-11-06 16:02:13,054 INFO mapreduce.Job: map 30% reduce 0%
2022-11-06 16:02:16,069 INFO mapreduce.Job: map 32% reduce 0%
2022-11-06 16:02:18,081 INFO mapreduce.Job: map 37% reduce 0%
2022-11-06 16:02:20,089 INFO mapreduce.Job: map 41% reduce 0%
2022-11-06 16:02:20,090 INFO mapreduce.Job: map 42% reduce 0%
2022-11-06 16:02:22,103 INFO mapreduce.Job: map 44% reduce 0%
2022-11-06 16:02:25,114 INFO mapreduce.Job: map 46% reduce 0%
2022-11-06 16:02:26,147 INFO mapreduce.Job: map 52% reduce 0%
2022-11-06 16:02:28,157 INFO mapreduce.Job: map 56% reduce 0%
2022-11-06 16:02:31,171 INFO mapreduce.Job: map 60% reduce 0%
2022-11-06 16:02:32,174 INFO mapreduce.Job: map 64% reduce 0%
2022-11-06 16:02:33,179 INFO mapreduce.Job: map 67% reduce 0%
2022-11-06 16:02:34,183 INFO mapreduce.Job: map 68% reduce 0%
2022-11-06 16:02:36,204 INFO mapreduce.Job: map 75% reduce 0%
2022-11-06 16:02:37,211 INFO mapreduce.Job: map 75% reduce 23%
2022-11-06 16:02:38,215 INFO mapreduce.Job: map 79% reduce 23%
2022-11-06 16:02:43,239 INFO mapreduce.Job: map 79% reduce 26%
2022-11-06 16:02:48,273 INFO mapreduce.Job: map 81% reduce 26%
2022-11-06 16:02:51,283 INFO mapreduce.Job: map 83% reduce 26%
2022-11-06 16:02:53,298 INFO mapreduce.Job: map 84% reduce 26%
2022-11-06 16:02:54,308 INFO mapreduce.Job: map 87% reduce 26%
2022-11-06 16:02:57,311 INFO mapreduce.Job: map 88% reduce 26%

```

c. Show output

```

hadoopuser@assignment1-ankale-main:~$ hdfs dfs -cat /access
00   66.249.66.194    14298
00   66.249.66.91     12232
00   66.249.66.92     4291
01   66.249.66.91     13874
01   66.249.66.91     13875
01   66.249.66.92     2024
02   66.249.66.91     11697
02   66.249.66.194    19345
02   17.58.182.43     898
03   23.181.169.3     14144
03   66.249.66.91     7914
03   17.58.182.43     958
04   66.249.66.194    19885
04   66.249.66.91     7917
04   66.249.66.91     1000
04   192.168.1.3      1981
05   66.249.66.194    18534
05   66.249.66.91     7935
05   23.181.169.3     1382
05   66.249.66.194    18033
06   66.249.66.91     9248
06   23.181.169.3     14148
07   66.249.66.194    12267
07   66.249.66.91     9116
07   66.249.66.91     1000
08   66.249.66.194    12944
08   66.249.66.91     18237
08   151.239.241.163   6256
09   66.249.66.194    14033
09   66.249.66.194    14148
09   151.239.241.163   9149
10   66.249.66.194    17292
10   66.249.66.91     13213
10   10.10.10.10       9817
11   66.249.66.194    8572
11   66.249.66.91     13631
11   151.239.241.163   8642
12   66.249.66.194    16966
12   151.239.241.163   16966
12   151.239.241.163   8564
13   66.249.66.194    18372
13   66.249.66.91     15166
13   151.239.241.163   7924
14   66.249.66.194    17249
14   66.249.66.91     17893
14   151.239.241.163   8786
15   66.249.66.194    18273
16   66.249.66.194    18272
15   151.239.241.163   8564
16   66.249.66.91     17849
16   66.249.66.194    17512
16   151.239.241.163   7187
17   66.249.66.194    17104
17   66.249.66.91     17107
17   151.239.241.163   8573
18   66.249.66.194    17531
18   66.249.66.91     16727
19   66.249.66.91     171
19   66.249.66.91     18911
19   66.249.66.194    18569
19   104.222.32.9     9978
20   66.249.66.194    11864
20   66.249.66.194    15729
20   66.249.66.92     5589
21   66.249.66.194    14075
21   66.249.66.91     13783
22   66.249.66.91     4079
22   66.249.66.91     14894
22   66.249.66.194    13576
22   66.249.66.92     4901
23   66.249.66.194    14395
23   66.249.66.194    14392
23   66.249.66.92     4259

```

Part 2

2.1 MapReduce like a query

Query the MapReduce run using command line arguments. User input 'x-y' will output top 3 IPs between x and y hours (24 hr format).

Only change from Part1 is the mapper which accepts command line inputs.

Example 1: arg = '3-4'

a. Run command

```
~ - hadoopuser@assignment1-ankale-main: ~ -- ssh exouser@149.165.170.253          ~Library/CloudStorage/OneDrive-IndianaUniversity/FA22/E516 - Cloud Computing/A1 -- zsh
hadoopuser@assignment1-ankale-main:~$ hadoop jar /home/hadoopuser/hadoop-3.3.4/share/hadoop/tools/lib/hadoop-streaming-3.3.4.jar -input /accessLog/access.log -output /accessLog/queryOutput -mapper "python3 /home/hadoopuser/part1/mapper-query.py '3-4'" -reducer "python3 /home/hadoopuser/part1/reducer-main.py"
packageJobJar: [/tmp/hadoop-unjar1293142674066315927/] [] /tmp/streamjob2879349076687822438.jar tmpDir=null
2022-11-06 20:38:02,536 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /127.0.0.1:8032
2022-11-06 20:38:02,662 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /127.0.0.1:8032
2022-11-06 20:38:02,836 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoopuser/.staging/job_1667431627519_0014
2022-11-06 20:38:03,020 INFO mapred.FileInputFormat: Total input files to process : 1
2022-11-06 20:38:03,037 INFO net.NetworkTopology: Adding a new node: /default-rack/127.0.0.1:9866
2022-11-06 20:38:03,086 INFO mapreduce.JobSubmitter: number of splits:26
2022-11-06 20:38:03,234 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1667431627519_0014
2022-11-06 20:38:03,234 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-11-06 20:38:03,351 INFO conf.Configuration: resource-types.xml not found
2022-11-06 20:38:03,351 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-11-06 20:38:03,395 INFO impl.YarnClientImpl: Submitted application application_1667431627519_0014
2022-11-06 20:38:03,420 INFO mapreduce.Job: The url to track the job: http://assignment1-ankale-main.js2local:8088/proxy/application_1667431627519_0014/
2022-11-06 20:38:03,421 INFO mapreduce.Job: Running job: job_1667431627519_0014
2022-11-06 20:38:08,489 INFO mapreduce.Job: Job job_1667431627519_0014 running in uber mode : false
2022-11-06 20:38:08,490 INFO mapreduce.Job: map 0% reduce 0%
2022-11-06 20:38:13,558 INFO mapreduce.Job: Task Id : attempt_1667431627519_0014_m_000004_0, Status : FAILED
Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2
        at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:326)
        at org.apache.hadoop.streaming.PipeMapRed.mapredFinished(PipeMapRed.java:539)
        at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)
        at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)
        at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)
        at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:466)
        at org.apache.hadoop.mapred.MapTask.run(MapTask.java:250)
```

b. Show output:

```
Merged Map Outputs: 20
GC time elapsed (ms)=304610
CPU time spent (ms)=859
Physical memory (bytes) snapshot=9522823168
Virtual memory (bytes) snapshot=73844961280
Total committed heap usage (bytes)=13419675648
Peak Map Physical memory (bytes)=427077632
Peak Map Virtual memory (bytes)=2771259392
Peak Reduce Physical memory (bytes)=281579520
Peak Reduce Virtual memory (bytes)=2740097024
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=3502543223
File Output Format Counters
  Bytes Written=63
2022-11-06 21:27:37,775 INFO streaming.StreamJob: Output directory: /accessLog/queryOutput
[hadoopuser@assignment1-ankale-main:~$ hdfs dfs -ls /accessLog/queryOutput
Found 2 items
-rw-r--r-- 1 hadoopuser supergroup         0 2022-11-06 21:27 /accessLog/queryOutput/_SUCCESS
-rw-r--r-- 1 hadoopuser supergroup       63 2022-11-06 21:27 /accessLog/queryOutput/part-00000
[hadoopuser@assignment1-ankale-main:~$ hdfs dfs -cat /accessLog/queryOutput/part-00000
03      66.249.66.194    8644
03      66.249.66.91     7914
03      17.58.102.43    950
hadoopuser@assignment1-ankale-main:~$ ]
```

Example 2: arg = '1-3'

a. Run command

```
~ hadoopuser@assignment1-ankale-main:~ ssh exouser@149.165.170.253 -/Library/CloudStorage/OneDrive-IndianaUniversity/FA22/E516 - Cloud Computing/A1 --zsh
hadoopuser@assignment1-ankale-main:~$ hadoop jar /home/hadoopuser/hadoop-3.3.4/share/hadoop/tools/lib/hadoop-streaming-3.3.4.jar -input /accessLog/access.log -output /accessLog/queryOutput2 -mapper "python3 /home/hadoopuser/part2/mapper-query.py 1-3" -reducer "python3 /home/hadoopuser/part2/reducer-main.py"
packageJobJar: [/tmp/hadoop-unjar8938041817613626584/] []
/tmp/streamjob10877398248410373825.jar tmpDir=null
2022-11-06 21:34:05.234 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /127.0.0.1:8032
2022-11-06 21:34:05.417 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /127.0.0.1:8032
2022-11-06 21:34:05.635 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoopuser/.staging/job_1667431627519_0017
2022-11-06 21:34:05.882 INFO mapred.FileInputFormat: Total input files to process : 1
2022-11-06 21:34:05.898 INFO net.NetworkTopology: Adding a new node: /default-rack/127.0.0.1:9866
2022-11-06 21:34:05.933 INFO mapreduce.JobSubmitter: number of splits:26
2022-11-06 21:34:06.117 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1667431627519_0017
2022-11-06 21:34:06.118 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-11-06 21:34:06.269 INFO conf.Configuration: resource-types.xml not found
2022-11-06 21:34:06.269 INFO mapreduce.ResourceUtils: Unable to find 'resource-types.xml'.
2022-11-06 21:34:06.324 INFO impl.YarnClientImpl: Submitted application application_1667431627519_0017
2022-11-06 21:34:06.349 INFO mapreduce.Job: The url to track the job: http://assignment1-ankale-main.js2local:8088/proxy/application_1667431627519_0017/
2022-11-06 21:34:06.358 INFO mapreduce.Job: Running job: job_1667431627519_0017
2022-11-06 21:35:08.831 INFO mapreduce.Job: Job job_1667431627519_0017 running in uber mode : false
2022-11-06 21:35:08.831 INFO mapreduce.Job: map 0% reduce 0%
2022-11-06 21:35:18.928 INFO mapreduce.Job: map 8% reduce 0%
2022-11-06 21:35:26.989 INFO mapreduce.Job: map 9% reduce 0%
2022-11-06 21:35:27.994 INFO mapreduce.Job: map 14% reduce 0%
2022-11-06 21:35:33.028 INFO mapreduce.Job: map 16% reduce 0%
2022-11-06 21:35:34.034 INFO mapreduce.Job: map 20% reduce 0%
2022-11-06 21:35:35.044 INFO mapreduce.Job: map 21% reduce 0%
2022-11-06 21:35:40.002 INFO mapreduce.Job: map 22% reduce 0%
```

b. Show output:

```
Peak Reduce Physical Memory (bytes)=302940160
Peak Reduce Virtual memory (bytes)=2743742464
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=3502543223
File Output Format Counters
Bytes Written=131
2022-11-06 21:36:47.569 INFO streaming.StreamJob: Output directory: /accessLog/queryOutput2
hadoopuser@assignment1-ankale-main:~$ hdfs dfs -cat /accessLog/queryOutput2/part-00000
01 66.249.66.91 13874
01 66.249.66.194 12485
01 66.249.66.92 2924
02 66.249.66.91 11697
02 66.249.66.194 10345
02 17.58.102.43 898
hadoopuser@assignment1-ankale-main:~$
```

2.2 Fair and capacity scheduler

- a. Capacity scheduler
 - i. Change configs for queues over capacity scheduler
 - ii. Run stop-all.sh and start-all.sh
 - iii. Refresh queues using
> yarn rmadmin -refreshQueues
 - iv. Run 3 jobs (wordcount on access.log, wordcount on sample.log and grep on sample.log) simultaneously:

```

Terminal Shell Edit View Window Help
fair_scheduler_configs
mapreduce/hadoop-mapreduce-examples-3.3.4.jar wordcount /access
ss/putcapfair2
2022-11-09 03:36:11,468 INFO client.DefaultNoHARMFalloverProxyP
ing to ResourceManager at /127.0.0.1:8032
2022-11-09 03:36:11,989 INFO mapreduce.JobResourceUploader: Dis
ting for path: /tmp/hadoop-yarn/staging/fair/.staging/job_16679
2022-11-09 03:36:12,700 INFO input.FileInputFormat: Total input
: 1
2022-11-09 03:36:13,223 INFO mapreduce.JobSubmitter: number of
2022-11-09 03:36:13,934 INFO mapreduce.JobSubmitter: Submitting
job_1667964938421_0001
2022-11-09 03:36:13,935 INFO mapreduce.JobSubmitter: Executing
2022-11-09 03:36:14,118 INFO conf.Configuration: resource-types
2022-11-09 03:36:14,118 INFO resource.ResourceUtils: Unable to find 'resource-t
ypes.xml'.
2022-11-09 03:36:14,472 INFO impl.YarnClientImpl: Submitted applicat
ion_1667964938421_0001
2022-11-09 03:36:14,499 INFO mapreduce.Job: The url to track the job: http://ass
ignment1-ankale-main.js2local:8088/proxy/application_1667964938421_0001/
2022-11-09 03:36:14,500 INFO mapreduce.Job: Running job: job_1/
2022-11-09 03:36:21,611 INFO mapreduce.Job: Job job_1667964938
n uber mode : false
2022-11-09 03:36:21,612 INFO mapreduce.Job: map 0% reduce 0%
2022-11-09 03:36:21,612 INFO mapreduce.Job: map 0% reduce 0%
2022-11-09 03:36:21,612 INFO mapreduce.Job: map 0% reduce 0%
2022-11-09 03:36:14,071 INFO conf.Configuration: resource-type
s.xml not found
2022-11-09 03:36:14,071 INFO resource.ResourceUtils: Unable to find 'resource-t
ypes.xml'.
2022-11-09 03:36:14,461 INFO impl.YarnClientImpl: Submitted applicat
ion_1667964938421_0002
2022-11-09 03:36:14,497 INFO mapreduce.Job: The url to track t
he job: http://assignment1-ankale-main.js2local:8088/proxy/app
lication_1667964938421_0002/
2022-11-09 03:36:14,498 INFO mapreduce.Job: Running job: job_1
667964938421_0002
2022-11-09 03:36:20,673 INFO mapreduce.Job: Job job_1667964938
421_0002 running in uber mode : false
2022-11-09 03:36:20,674 INFO mapreduce.Job: map 0% reduce 0%
2022-11-09 03:36:25,735 INFO mapreduce.Job: map 100% reduce 0%

```

V. Schedules:

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU vCores	Allocated Memory MB	All
application_166796128057_0007	fair	grep-search	MAPREDUCE		default	0	Tue Nov 8 22:20:30 -0500 2022	N/A	N/A	ACCEPTED	UNDEFINED	0	0	0	-1
application_166796128057_0006	fair	word count	MAPREDUCE		default	0	Tue Nov 8 22:20:30 -0500 2022	N/A	N/A	ACCEPTED	UNDEFINED	0	0	0	-1
application_166796128057_0005	fair	word count	MAPREDUCE		default	0	Tue Nov 8 22:20:30 -0500 2022	Tue Nov 8 22:20:30 -0500 2022	N/A	RUNNING	UNDEFINED	7	7	8192	-1

Showing 1 to 3 of 3 entries

Aggregate scheduler counts

Total Container Allocations(count)	Total Container Releases(count)	Total Fulfilled Reservations(count)
21	14	0

Last scheduler run Time

Wed Nov 09 03:20:36 +0000 2022

Allocations(count - resources)

0 - <memory 0, vCores 0>

Last Preemption Time

Container Id

N/A

Last Reservation N/A

b. Fair share scheduler:

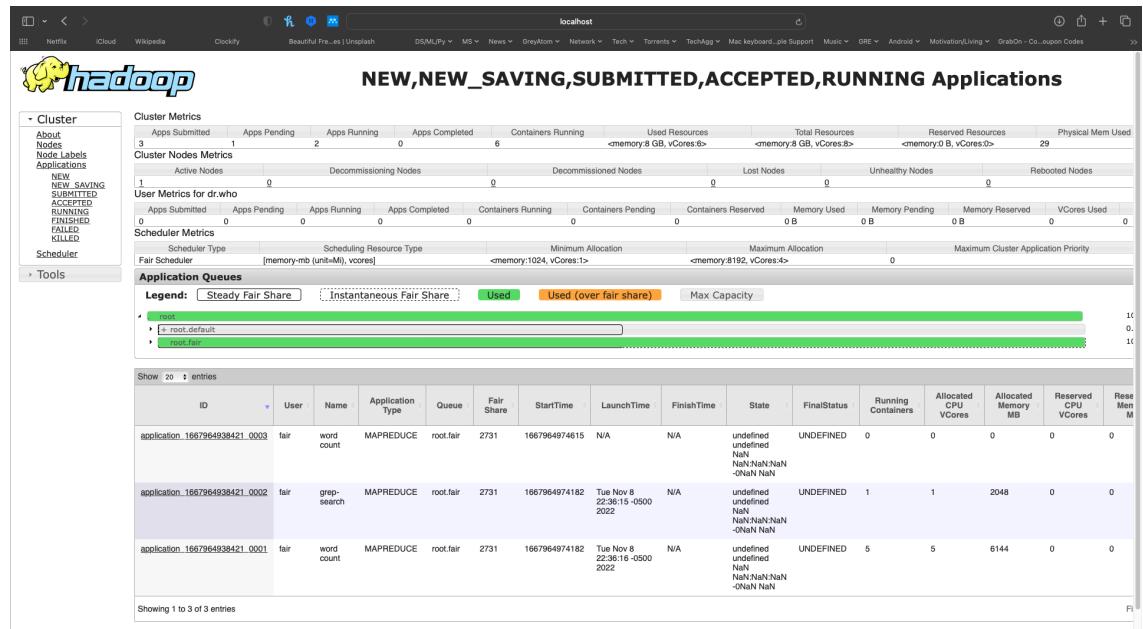
- Change yarn-site.xml config
- Run stop-all.sh and start-all.sh and check for “fair share” in localhost:8088 port
- Run 3 commands (wordcount on access.log, wordcount on sample.log and grep on sample.log) simultaneously:

```

split1:1
2022-11-09 03:20:31,505 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_16679612268057_0007
2022-11-09 03:20:31,505 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-11-09 03:20:31,505 INFO mapreduce.JobSubmitter: Configuration: resource-type
2022-11-09 03:20:31,505 INFO mapreduce.JobSubmitter: resource-type
2022-11-09 03:20:31,505 INFO mapreduce.JobSubmitter: Unable to find 'resource-types.xml'.
2022-11-09 03:20:31,505 INFO mapreduce.JobSubmitter: Submitted application application_16679612268057_0007
2022-11-09 03:20:31,505 INFO mapreduce.JobSubmitter: Submitted job: http://assignment1-ankale-main.js2local:8088/proxy/app
2022-11-09 03:20:31,505 INFO mapreduce.JobSubmitter: Exclamation_16679612268057_0007/
2022-11-09 03:20:31,505 INFO mapreduce.JobSubmitter: Running job: job_16679612268057_0007
2022-11-09 03:20:31,505 INFO mapreduce.JobSubmitter: Un
2022-11-09 03:20:29,209 INFO mapreduce.JobSubmitter: Application application_16679612268057_0005
2022-11-09 03:20:29,448 INFO mapreduce.JobSubmitter: Submitted application application_16679612268057_0005
2022-11-09 03:20:29,448 INFO mapreduce.JobSubmitter: Submitted job: http://assignment1-ankale-main.js2local:8088/proxy/application_16679612268057_0005
2022-11-09 03:20:29,448 INFO mapreduce.JobSubmitter: Exclamation_16679612268057_0005/
2022-11-09 03:20:29,637 INFO mapreduce.JobSubmitter: Running job: job_16679612268057_0005
2022-11-09 03:20:29,637 INFO mapreduce.JobSubmitter: Un
2022-11-09 03:20:29,723 INFO mapreduce.JobSubmitter: Submitted application application_16679612268057_0005
2022-11-09 03:20:29,758 INFO mapreduce.JobSubmitter: Job: The url to track the job: http://assignment1-ankale-main.js2local:8088/proxy/application_16679612268057_0005/
2022-11-09 03:20:29,758 INFO mapreduce.JobSubmitter: Running job: job_16679612268057_0005
2022-11-09 03:20:35,840 INFO mapreduce.JobSubmitter: Job job_16679612268057_0005 running in uber mode : false
2022-11-09 03:20:35,842 INFO mapreduce.JobSubmitter: map 0% reduce 0%

```

iv. Schedules:



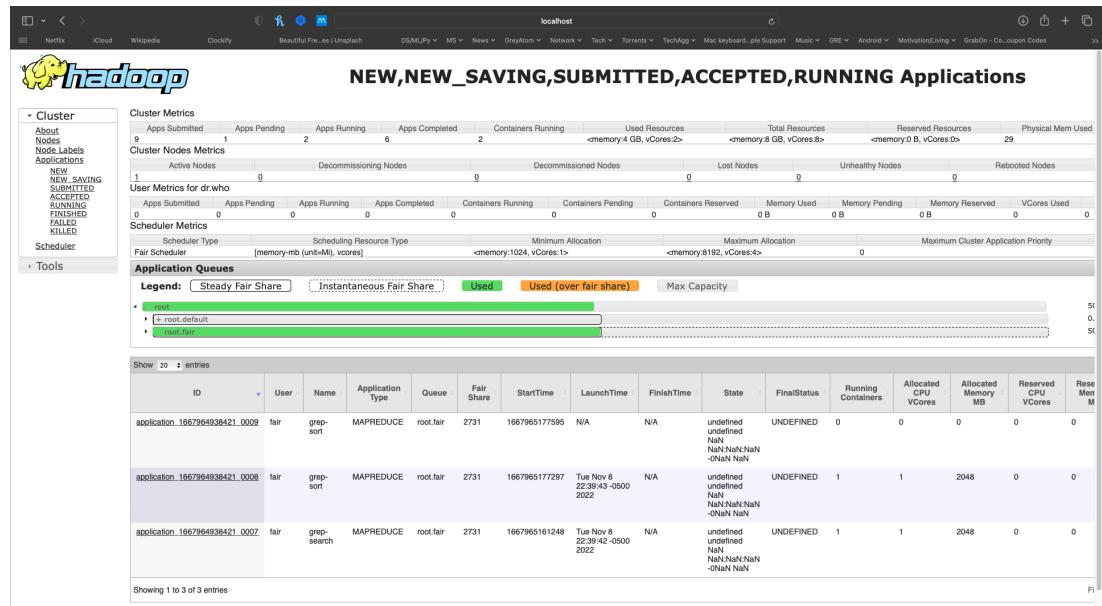
V. Another experiment: Run 3 jobs (grep on sample.log x3) simultaneously

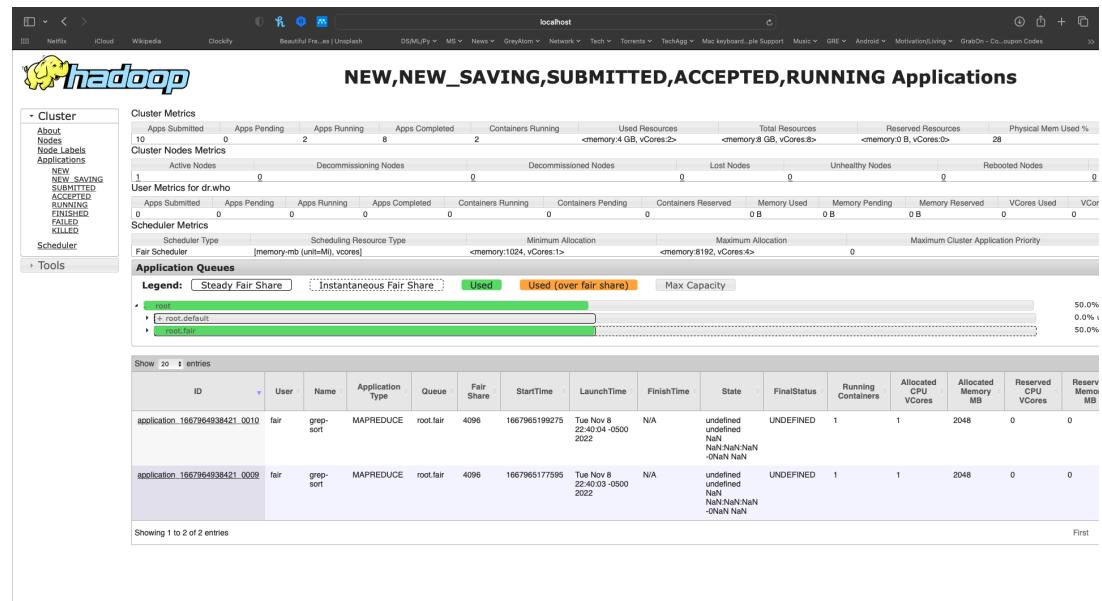
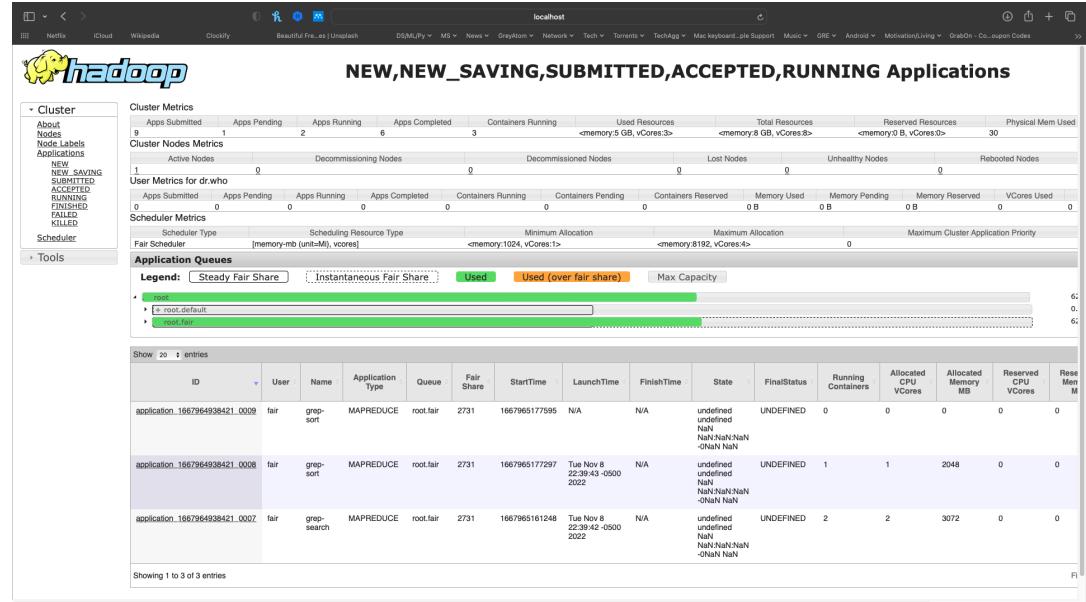
```

IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=102399
File Output Format Counters
Bytes Written=104
2022-11-09 03:39:37,456 INFO client.DefaultNoHARMFailoverProxyP 2022-11-09 03:39:47,510 INFO mapreduce.Job: map 0% redu
ng to ResourceManager at /127.0.0.1:18032
2022-11-09 03:39:37,476 INFO mapreduce.JobResourceUploader: Dis 2022-11-09 03:39:52,576 INFO mapreduce.Job: map 100% re
ding for path: /tmp/hadoop-yarn/staging/fair-/staging/job_16679
2022-11-09 03:39:37,517 INFO input.FileInputFormat: Total input
: 1
2022-11-09 03:39:37,549 INFO mapreduce.JobSubmitter: number of splits:1
2022-11-09 03:39:37,580 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_1667964938421_0009
2022-11-09 03:39:37,580 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-11-09 03:39:37,601 INFO impl.YarnClientImpl: Submitted application applicat
ion_1667964938421_0009
2022-11-09 03:39:37,605 INFO mapreduce.Job: The url to track the job: http://ass
ignment1-ankale-main.js2local:8088/proxy/application_1667964938421_0009/
2022-11-09 03:39:37,605 INFO mapreduce.Job: Running job: job_1667964938421_0009
2022-11-09 03:39:37,301 INFO impl.YarnClientImpl: Submitted ap
plication application_1667964938421_0008
2022-11-09 03:39:37,307 INFO mapreduce.Job: The url to track t
he job: http://assignment1-ankale-main.js2local:8088/app
lication_1667964938421_0008/
2022-11-09 03:39:37,307 INFO mapreduce.Job: Running job: job_1
667964938421_0008
2022-11-09 03:39:48,416 INFO mapreduce.Job: Job job_1667964938421_0008
running in uber mode : false
2022-11-09 03:39:48,417 INFO mapreduce.Job: map 0% reduce 0%
2022-11-09 03:39:53,458 INFO mapreduce.Job: map 100% reduce 0%

```

vi. Schedules





Comparison of Capacity v/s Fair scheduler:

The Capacity scheduler schedules jobs in queues. But the fair scheduler gives priority to the first job then allocates resources to the second and then third job according to the requirements. In the fair share example 1, the highest priority was given to the most resource intensive job, followed by the other. This can also be seen in example 2 for fair share, where resources were almost equally shared for scheduled jobs of the same type (i.e. grep).

The Jobs run:

Wordcount 1 on sample.log

hadoop jar

/home/fair/hadoop-3.3.4/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.4.jar

wordcount /sample/sample.log /sample/output

Wordcount 1 on access.log

hadoop jar

/home/fair/hadoop-3.3.4/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.4.jar

wordcount /access/access.log /access/output

Grep on sample.log

hadoop jar

/home/fair/hadoop-3.3.4/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.4.jar grep

/sample/sample.log /sample/outputGrepfair33 '.'*

Bonus:

I have run all my codes in the Jetstream instance and the access.log files takes less than 2 minutes to complete. Refer screenshots above.