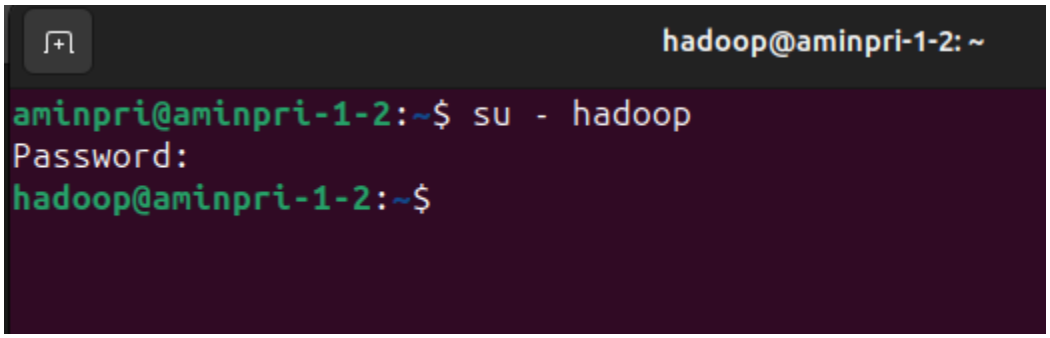
Prithvi Amin (aminpri)
SP24-BL-ENGR-E516-14573

ASSIGNMENT 1

1. Describe your Jetstream2 instance. What was your cloud.init script? Which size instance did you use?
2. Did you use the Console, Web Shell or Web Desktop? If you used more than one interface, which did you prefer?
3. Do you have any feedback on your experience with this instance and interface(s)?
    - **I didn't use the Jetstream instance. I worked on my VMware for this assignment.**

Switch from main user to hadoop user:



Start hadoop cluster:

```
hadoop@aminpri-1-2:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [aminpri-1-2]
Starting resourcemanager
Starting nodemanagers
hadoop@aminpri-1-2:~$
```

Check status:

```
hadoop@aminpri-1-2:~$ jps
6048 SecondaryNameNode
6417 NodeManager
6835 Jps
5796 DataNode
5627 NameNode
6287 ResourceManager
hadoop@aminpri-1-2:~$
```

Hadoop cluster is up and running.

Part1:

Code for driver.py:

```
hadoop@aminpri-1-2:~/assignment1/Part_01$ cat driver.py
#!/usr/bin/env python

import subprocess

input_file = "../Input/access.log"
output_file = "/"

# Run the first MapReduce job to count occurrences of IP addresses per hour
mapper1_cmd = ["python3", "mapper.py"]
reducer1_cmd = ["python3", "reducer.py"]
sort_cmd = ["sort"]

with open(input_file, 'r') as f:
    p1 = subprocess.Popen(mapper1_cmd, stdin=f, stdout=subprocess.PIPE)
    p2 = subprocess.Popen(sort_cmd, stdin=p1.stdout, stdout=subprocess.PIPE)
    p3 = subprocess.Popen(reducer1_cmd, stdin=p2.stdout)

p3.wait()
hadoop@aminpri-1-2:~/assignment1/Part_01$
```

Code for mapper.py:

```
hadoop@aminpri-1-2:~/assignment1/Part_01$ cat mapper.py
import sys
import re
from datetime import datetime

# Regular expression pattern to match IP addresses and timestamps
log_pattern = re.compile(r'(?P<ip>\d+\.\d+\.\d+\.\d+) .* \[(?P<timestamp>[^\]]+)\]')

for line in sys.stdin:
    match = log_pattern.search(line.strip())
    if match:
        ip = match.group('ip')
        timestamp_str = match.group('timestamp').split()[0]  # Extracting date from timestamp
        try:
            timestamp = datetime.strptime(timestamp_str, "%d/%b/%Y:%H:%M:%S")
            hour = timestamp.strftime("%H")
            print(f"{hour}\t{ip}\t1")  # Emitting hour, IP, and count
        except ValueError:
            pass
hadoop@aminpri-1-2:~/assignment1/Part_01$
```

Code for reducer.py:

```
hadoop@aminpri-1-2:~/assignment1/Part_01$ cat reducer.py
#!/usr/bin/env python3

import sys
from operator import itemgetter
from collections import defaultdict
from datetime import datetime, timedelta

hourly_ip_count = defaultdict(lambda: defaultdict(int))
total_ip_count = defaultdict(int)

for line in sys.stdin:
    line = line.strip()
    hour, ip, count = line.split('\t', 2)
    count = int(count)
    hourly_ip_count[hour][ip] += count
    total_ip_count[ip] += count

# Sort IP addresses based on their total counts
sorted_total_ips = sorted(total_ip_count.items(), key=itemgetter(1), reverse=True)
```

```
# Output the top 3 IP addresses in the dataset
print("Top 3 IP addresses in the dataset:")
for ip, count in sorted_total_ips[:3]:
    print(f"IP: {ip}, Count: {count}")

# Iterate through each hour
print("\nTop 3 IP addresses with the granularity of an hour:")
for hour, ip_counts in hourly_ip_count.items():
    start_time = datetime.strptime(hour, "%H")
    end_time = start_time + timedelta(hours=1) - timedelta(seconds=1)
    hour_range = f"{start_time.strftime('%H:%M:%S')} to {end_time.strftime('%H:%M:%S')}"
    # Sort IP addresses based on their counts for this hour
    sorted_ips = sorted(ip_counts.items(), key=itemgetter(1), reverse=True)
    # Output the top 3 IP addresses for each hour
    print(f"\nFrom hour {hour_range}:")
    for ip, count in sorted_ips[:3]:
        print(f"IP: {ip}, Count: {count}")
hadoop@aminpri-1-2:~/assignment1/Part_01$
```

Place the files in the correct directories:

```
hadoop@aminpri-1-2:~$ ls
assignment1  hadoop  hadoop-3.3.6.tar.gz  hadoopdata  snap
hadoop@aminpri-1-2:~$ cd assignment1
hadoop@aminpri-1-2:~/assignment1$ ls
Input  Part_01  Part_02
hadoop@aminpri-1-2:~/assignment1$ cd Input
hadoop@aminpri-1-2:~/assignment1/Input$ ls
access.log  access.log.zip  sample.log  sample.txt
hadoop@aminpri-1-2:~/assignment1/Input$ cd ../Part_01
hadoop@aminpri-1-2:~/assignment1/Part_01$ ls
driver.py  mapper.py  reducer.py
hadoop@aminpri-1-2:~/assignment1/Part_01$
```

Mount to hdfs:

```
hadoop@aminpri-1-2:~$ hdfs dfs -copyFromLocal "assignment1/" /
hadoop@aminpri-1-2:~$
```

Check in browser: http://localhost:9870/

/assignment1/Part_01 Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	523 B	Mar 09 15:08	1	128 MB	driver.py
-rw-r--r--	hadoop	supergroup	660 B	Mar 09 15:08	1	128 MB	mapper.py
-rw-r--r--	hadoop	supergroup	1.34 KB	Mar 09 15:08	1	128 MB	reducer.py

/assignment1/Input Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	3.26 GB	Mar 09 15:08	1	128 MB	access.log
-rw-r--r--	hadoop	supergroup	264.27 MB	Mar 09 15:08	1	128 MB	access.log.zip
-rw-r--r--	hadoop	supergroup	100 KB	Mar 09 15:08	1	128 MB	sample.log
-rw-r--r--	hadoop	supergroup	100 KB	Mar 09 15:08	1	128 MB	sample.txt

Run the below command to start the job:

hadoop jar /$HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar -file /home/hadoop/assignment1/Part_01/mapper.py -mapper "/usr/bin/python3 /home/hadoop/assignment1/Part_01/mapper.py" -file /home/hadoop/assignment1/Part_01/reducer.py -reducer "/usr/bin/python3 /home/hadoop/assignment1/Part_01/reducer.py" -input /assignment1/Input/access.log -output /assignment1/Output_part1

```
hadoop@aminpri-1-2:~$ hadoop jar /$HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar -file /home/hadoop
/assignment1/Part_01/mapper.py -mapper "/usr/bin/python3 /home/hadoop/assignment1/Part_01/mapper.py" -file /hom
e/hadoop/assignment1/Part_01/reducer.py -reducer "/usr/bin/python3 /home/hadoop/assignment1/Part_01/reducer.py"
 -input /assignment1/Input/access.log -output /assignment1/Output_part1
2024-03-09 15:19:59,096 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/hadoop/assignment1/Part_01/mapper.py, /home/hadoop/assignment1/Part_01/reducer.py, /tmp/hadoop-un
jar8019172214574102786/] [] /tmp/streamjob5606847591153862966.jar tmpDir=null
2024-03-09 15:20:00,581 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-03-09 15:20:00,964 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-03-09 15:20:01,338 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging
/hadoop/.staging/job_1710011891143_0002
2024-03-09 15:20:01,797 INFO mapred.FileInputFormat: Total input files to process : 1
2024-03-09 15:20:01,847 INFO net.NetworkTopology: Adding a new node: /default-rack/127.0.0.1:9866
2024-03-09 15:20:01,916 INFO mapreduce.JobSubmitter: number of splits:26
2024-03-09 15:20:02,219 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1710011891143_0002
2024-03-09 15:20:02,219 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-03-09 15:20:02,482 INFO conf.Configuration: resource-types.xml not found
2024-03-09 15:20:02,483 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-03-09 15:20:02,593 INFO impl.YarnClientImpl: Submitted application application_1710011891143_0002
2024-03-09 15:20:02,665 INFO mapreduce.Job: The url to track the job: http://aminpri-1-2:8088/proxy/application_1710011
891143_0002/
2024-03-09 15:20:02,667 INFO mapreduce.Job: Running job: job_1710011891143_0002
2024-03-09 15:20:09,792 INFO mapreduce.Job: Job job_1710011891143_0002 running in uber mode : false
```

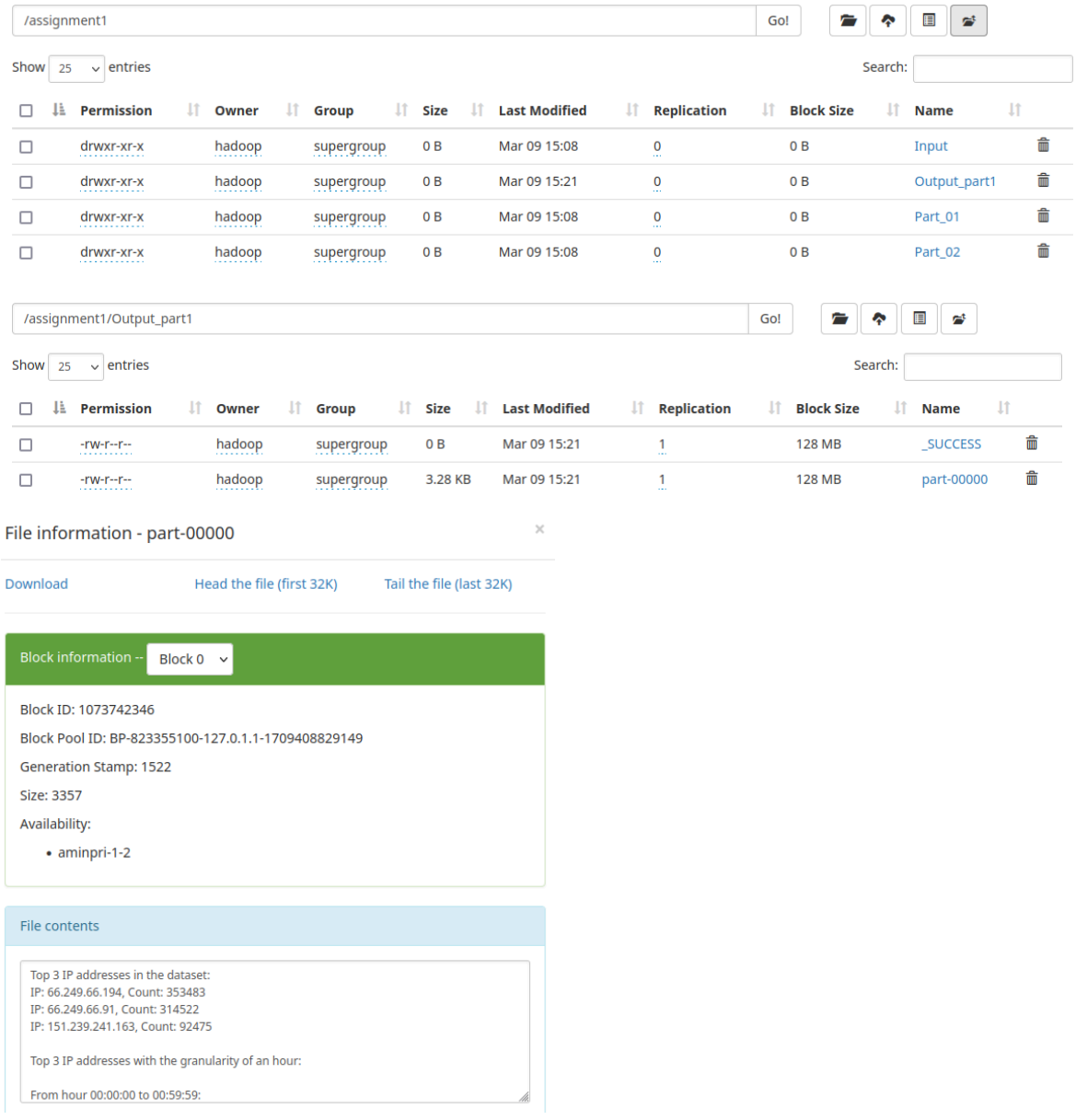
```
2024-03-09 15:20:09,792 INFO mapreduce.Job:  map 0% reduce 0%
2024-03-09 15:20:27,201 INFO mapreduce.Job:  map 3% reduce 0%
2024-03-09 15:20:28,310 INFO mapreduce.Job:  map 12% reduce 0%
2024-03-09 15:20:29,327 INFO mapreduce.Job:  map 18% reduce 0%
2024-03-09 15:20:30,350 INFO mapreduce.Job:  map 20% reduce 0%
2024-03-09 15:20:31,371 INFO mapreduce.Job:  map 21% reduce 0%
2024-03-09 15:20:32,388 INFO mapreduce.Job:  map 23% reduce 0%
2024-03-09 15:20:46,680 INFO mapreduce.Job:  map 25% reduce 0%
2024-03-09 15:20:47,689 INFO mapreduce.Job:  map 27% reduce 0%
2024-03-09 15:20:48,706 INFO mapreduce.Job:  map 29% reduce 0%
2024-03-09 15:20:49,939 INFO mapreduce.Job:  map 31% reduce 0%
2024-03-09 15:20:50,952 INFO mapreduce.Job:  map 39% reduce 0%
2024-03-09 15:20:52,984 INFO mapreduce.Job:  map 44% reduce 0%
2024-03-09 15:20:53,991 INFO mapreduce.Job:  map 46% reduce 0%
2024-03-09 15:21:08,377 INFO mapreduce.Job:  map 54% reduce 0%
2024-03-09 15:21:09,415 INFO mapreduce.Job:  map 54% reduce 18%
2024-03-09 15:21:10,465 INFO mapreduce.Job:  map 58% reduce 18%
2024-03-09 15:21:11,497 INFO mapreduce.Job:  map 60% reduce 18%
2024-03-09 15:21:12,579 INFO mapreduce.Job:  map 64% reduce 18%
2024-03-09 15:21:14,620 INFO mapreduce.Job:  map 65% reduce 18%
2024-03-09 15:21:15,634 INFO mapreduce.Job:  map 65% reduce 21%
2024-03-09 15:21:21,736 INFO mapreduce.Job:  map 65% reduce 22%
2024-03-09 15:21:25,866 INFO mapreduce.Job:  map 68% reduce 22%
2024-03-09 15:21:27,888 INFO mapreduce.Job:  map 74% reduce 22%
2024-03-09 15:21:28,903 INFO mapreduce.Job:  map 75% reduce 22%
2024-03-09 15:21:29,910 INFO mapreduce.Job:  map 81% reduce 22%
2024-03-09 15:21:32,935 INFO mapreduce.Job:  map 83% reduce 22%
2024-03-09 15:21:33,951 INFO mapreduce.Job:  map 83% reduce 27%
2024-03-09 15:21:34,957 INFO mapreduce.Job:  map 85% reduce 27%
2024-03-09 15:21:40,013 INFO mapreduce.Job:  map 85% reduce 28%
2024-03-09 15:21:46,138 INFO mapreduce.Job:  map 96% reduce 28%
2024-03-09 15:21:48,002 INFO mapreduce.Job:  map 100% reduce 28%
2024-03-09 15:21:52,025 INFO mapreduce.Job:  map 100% reduce 72%
```

```
2024-03-09 15:22:00,099 INFO mapreduce.Job: Job job_1710011891143_0002 completed successfully
2024-03-09 15:22:00,251 INFO mapreduce.Job: Counters: 56
        File System Counters
                FILE: Number of bytes read=214353104
                FILE: Number of bytes written=436263295
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3502545875
                HDFS: Number of bytes written=3357
                HDFS: Number of read operations=83
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=27
                Launched reduce tasks=1
                Data-local map tasks=26
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=487381
                Total time spent by all reduces in occupied slots (ms)=67599
                Total time spent by all map tasks (ms)=487381
                Total time spent by all reduce tasks (ms)=67599
                Total vcore-milliseconds taken by all map tasks=487381
                Total vcore-milliseconds taken by all reduce tasks=67599
                Total megabyte-milliseconds taken by all map tasks=499078144
                Total megabyte-milliseconds taken by all reduce tasks=69221376
```

```
                Peak Reduce Virtual memory (bytes)=2613448704
        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=3357
2024-03-09 15:22:00,251 INFO streaming.StreamJob: Output directory: /assignment1/Output_part1
hadoop@aminpri-1-2:~$
```

Job is completed successfully and the output file is created.

Check the output file on the browser:



Output file:
Top 3 IP addresses in the dataset:
IP: 66.249.66.194, Count: 353483
IP: 66.249.66.91, Count: 314522
IP: 151.239.241.163, Count: 92475

Top 3 IP addresses with the granularity of an hour:

From hour 00:00:00 to 00:59:59:
IP: 66.249.66.194, Count: 14298
IP: 66.249.66.91, Count: 12232
IP: 66.249.66.92, Count: 4291

From hour 01:00:00 to 01:59:59:
IP: 66.249.66.91, Count: 13874
IP: 66.249.66.194, Count: 12485
IP: 66.249.66.92, Count: 2924

From hour 02:00:00 to 02:59:59:
IP: 66.249.66.91, Count: 11697
IP: 66.249.66.194, Count: 10345
IP: 91.99.72.15, Count: 1448

From hour 03:00:00 to 03:59:59:
IP: 66.249.66.194, Count: 8644
IP: 66.249.66.91, Count: 7914
IP: 91.99.72.15, Count: 1275

From hour 04:00:00 to 04:59:59:
IP: 66.249.66.194, Count: 10805
IP: 66.249.66.91, Count: 7571
IP: 91.99.72.15, Count: 1511

From hour 05:00:00 to 05:59:59:
IP: 66.249.66.194, Count: 10534
IP: 66.249.66.91, Count: 7035
IP: 91.99.72.15, Count: 1921

From hour 06:00:00 to 06:59:59:
IP: 66.249.66.194, Count: 10283
IP: 66.249.66.91, Count: 7968
IP: 91.99.72.15, Count: 2051

From hour 07:00:00 to 07:59:59:

IP: 66.249.66.194, Count: 12267
IP: 66.249.66.91, Count: 9116
IP: 91.99.72.15, Count: 2295

From hour 08:00:00 to 08:59:59:
IP: 66.249.66.194, Count: 12964
IP: 66.249.66.91, Count: 10237
IP: 151.239.241.163, Count: 6256

From hour 09:00:00 to 09:59:59:
IP: 66.249.66.194, Count: 14833
IP: 66.249.66.91, Count: 11450
IP: 151.239.241.163, Count: 9169

From hour 10:00:00 to 10:59:59:
IP: 66.249.66.194, Count: 17292
IP: 66.249.66.91, Count: 13213
IP: 151.239.241.163, Count: 9824

From hour 11:00:00 to 11:59:59:
IP: 66.249.66.194, Count: 15572
IP: 66.249.66.91, Count: 13631
IP: 151.239.241.163, Count: 8642

From hour 12:00:00 to 12:59:59:
IP: 66.249.66.194, Count: 16966
IP: 66.249.66.91, Count: 12656
IP: 151.239.241.163, Count: 8564

From hour 13:00:00 to 13:59:59:
IP: 66.249.66.194, Count: 18372
IP: 66.249.66.91, Count: 16166
IP: 151.239.241.163, Count: 7801

From hour 14:00:00 to 14:59:59:
IP: 66.249.66.194, Count: 19249
IP: 66.249.66.91, Count: 17893
IP: 151.239.241.163, Count: 8786

From hour 15:00:00 to 15:59:59:
IP: 66.249.66.194, Count: 18273
IP: 66.249.66.91, Count: 16662
IP: 151.239.241.163, Count: 6558

From hour 16:00:00 to 16:59:59:
IP: 66.249.66.91, Count: 17849
IP: 66.249.66.194, Count: 17512
IP: 151.239.241.163, Count: 7187

From hour 17:00:00 to 17:59:59:
IP: 66.249.66.194, Count: 18954
IP: 66.249.66.91, Count: 17107
IP: 151.239.241.163, Count: 8571

From hour 18:00:00 to 18:59:59:
IP: 66.249.66.194, Count: 17531
IP: 66.249.66.91, Count: 16727
IP: 104.222.32.91, Count: 7159

From hour 19:00:00 to 19:59:59:
IP: 66.249.66.91, Count: 18911
IP: 66.249.66.194, Count: 18569
IP: 104.222.32.91, Count: 9076

From hour 20:00:00 to 20:59:59:
IP: 66.249.66.91, Count: 15834
IP: 66.249.66.194, Count: 15729
IP: 66.249.66.92, Count: 5589

From hour 21:00:00 to 21:59:59:
IP: 66.249.66.194, Count: 14075
IP: 66.249.66.91, Count: 13783
IP: 66.249.66.92, Count: 4552

From hour 22:00:00 to 22:59:59:
IP: 66.249.66.91, Count: 14094
IP: 66.249.66.194, Count: 13576
IP: 66.249.66.92, Count: 4901

From hour 23:00:00 to 23:59:59:
IP: 66.249.66.194, Count: 14355
IP: 66.249.66.91, Count: 10902
IP: 66.249.66.92, Count: 4259

Part2:

Code for driver.py:

```
hadoop@aminpri-1-2:~/assignment1/Part_02$ cat driver.py
#!/usr/bin/env python

import subprocess

input_file = "../Input/access.log"
output_file = "/"

# Run the first MapReduce job to count occurrences of IP addresses per hour
mapper1_cmd = ["python3", "mapper.py", ]
reducer1_cmd = ["python3", "reducer.py", "1-2"]
sort_cmd = ["sort"]

with open(input_file, 'r') as f:
    p1 = subprocess.Popen(mapper1_cmd, stdin=f, stdout=subprocess.PIPE)
    p2 = subprocess.Popen(sort_cmd, stdin=p1.stdout, stdout=subprocess.PIPE)
    p3 = subprocess.Popen(reducer1_cmd, stdin=p2.stdout)

p3.wait()
hadoop@aminpri-1-2:~/assignment1/Part_02$
```

Code for mapper.py:

```
hadoop@aminpri-1-2:~/assignment1/Part_02$ cat mapper.py
import sys
import re
from datetime import datetime

# Regular expression pattern to match IP addresses and timestamps
log_pattern = re.compile(r'(?P<ip>\d+\.\d+\.\d+\.\d+) .* \[(?P<timestamp>[^\]]+)\]')

for line in sys.stdin:
    match = log_pattern.search(line.strip())
    if match:
        ip = match.group('ip')
        timestamp_str = match.group('timestamp').split()[0]  # Extracting date from timestamp
        try:
            timestamp = datetime.strptime(timestamp_str, "%d/%b/%Y:%H:%M:%S")
            hour = timestamp.strftime("%Y-%m-%d %H")
            print(f"{hour}\t{ip}\t1")  # Emitting hour, IP, and count
        except ValueError:
            pass
hadoop@aminpri-1-2:~/assignment1/Part_02$
```

Code for reducer.py:

```
hadoop@aminpri-1-2:~/assignment1/Part_02$ cat reducer.py
import sys
from collections import defaultdict
from operator import itemgetter

hourly_ip_count = defaultdict(int)

# Accept user input for hour range
start_hour, end_hour = map(int, sys.argv[1].split('-'))

for line in sys.stdin:
    line = line.strip()
    hour, ip, count = line.split('\t', 2)
    count = int(count)

    # Extract the hour from the timestamp
    hour_int = int(hour[-2:])

    # Check if the hour falls within the user-defined range
    if start_hour <= hour_int < end_hour:
        hourly_ip_count[ip] += count

# Sort IP addresses based on their counts
sorted_ips = sorted(hourly_ip_count.items(), key=itemgetter(1), reverse=True)[:3]

# Output the top 3 IP addresses for the specified hour range
print(f"Top 3 IP addresses from {start_hour:02}:00:00 to {(end_hour-1):02}:59:59:")
for ip, count in sorted_ips:
    print(f"    IP: {ip}, Count: {count}")
hadoop@aminpri-1-2:~/assignment1/Part_02$
```

Running on local:

```
hadoop@aminpri-1-2:~/assignment1/Part_02$ python3 driver.py 1-2
Top 3 IP addresses from 01:00:00 to 01:59:59:
    IP: 66.249.66.91, Count: 13874
    IP: 66.249.66.194, Count: 12485
    IP: 66.249.66.92, Count: 2924
```

Mounted on hdfs:

/assignment1/Part_02 Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	532 B	Mar 09 15:07	1	128 MB	driver.py
-rw-r--r--	hadoop	supergroup	1 KB	Mar 09 15:08	1	128 MB	hadoop_job.sh
-rw-r--r--	hadoop	supergroup	669 B	Mar 09 15:07	1	128 MB	mapper.py
-rw-r--r--	hadoop	supergroup	880 B	Mar 09 15:07	1	128 MB	reducer.py

Run the below command to start the job:

hadoop jar /$HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar -file /home/hadoop/assignment1/Part_02/mapper.py -mapper "/usr/bin/python3

/home/hadoop/assignment1/Part_02/mapper.py" -file /home/hadoop/assignment1/Part_02/reducer.py -reducer "/usr/bin/python3 /home/hadoop/assignment1/Part_02/reducer.py 1-2" -input /assignment1/Input/access.log -output /assignment1/Output_part2_1_2

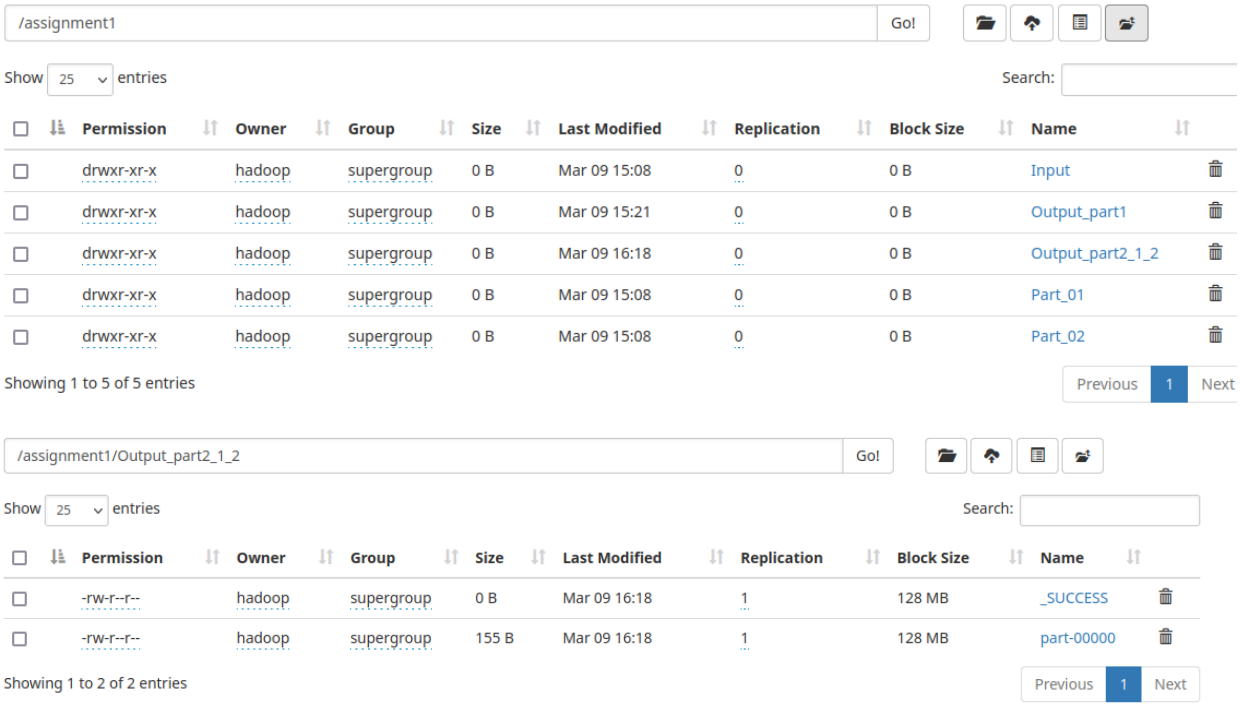
```
hadoop@aminpri-1-2:~$ hadoop jar /$HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar     -file /home/hadoop
/assignment1/Part_02/mapper.py     -mapper "/usr/bin/python3 /home/hadoop/assignment1/Part_02/mapper.py"     -file /hom
e/hadoop/assignment1/Part_02/reducer.py  -reducer "/usr/bin/python3 /home/hadoop/assignment1/Part_02/reducer.py 1-2"
 -input /assignment1/Input/access.log     -output /assignment1/Output_part2_1_2
2024-03-09 16:16:29,120 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/hadoop/assignment1/Part_02/mapper.py, /home/hadoop/assignment1/Part_02/reducer.py, /tmp/hadoop-un
jar6130701807954456760/] [] /tmp/streamjob7867603323123598849.jar tmpDir=null
2024-03-09 16:16:30,358 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-03-09 16:16:30,715 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-03-09 16:16:31,126 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging
/hadoop/.staging/job_1710011891143_0003
2024-03-09 16:16:32,568 INFO mapred.FileInputFormat: Total input files to process : 1
2024-03-09 16:16:32,611 INFO net.NetworkTopology: Adding a new node: /default-rack/127.0.0.1:9866
2024-03-09 16:16:32,744 INFO mapreduce.JobSubmitter: number of splits:26
2024-03-09 16:16:32,958 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1710011891143_0003
2024-03-09 16:16:32,958 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-03-09 16:16:33,263 INFO conf.Configuration: resource-types.xml not found
2024-03-09 16:16:33,263 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-03-09 16:16:33,420 INFO impl.YarnClientImpl: Submitted application application_1710011891143_0003
2024-03-09 16:16:33,515 INFO mapreduce.Job: The url to track the job: http://aminpri-1-2:8088/proxy/application_1710011
891143_0003/
2024-03-09 16:16:33,517 INFO mapreduce.Job: Running job: job_1710011891143_0003
```

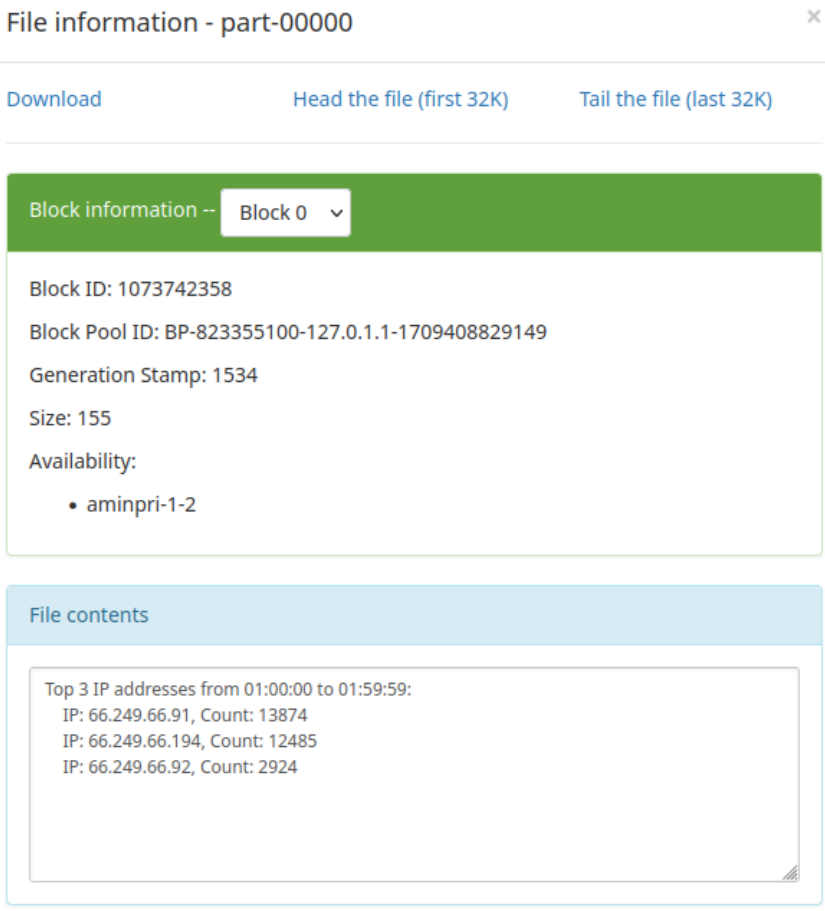
```
2024-03-09 16:18:29,316 INFO mapreduce.Job:  map 100% reduce 28%
2024-03-09 16:18:31,332 INFO mapreduce.Job:  map 100% reduce 59%
2024-03-09 16:18:37,373 INFO mapreduce.Job:  map 100% reduce 94%
2024-03-09 16:18:39,395 INFO mapreduce.Job:  map 100% reduce 100%
2024-03-09 16:18:39,402 INFO mapreduce.Job: Job job_1710011891143_0003 completed successfully
2024-03-09 16:18:39,524 INFO mapreduce.Job: Counters: 56
        File System Counters
                FILE: Number of bytes read=328369270
                FILE: Number of bytes written=664295816
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3502545875
                HDFS: Number of bytes written=155
                HDFS: Number of read operations=83
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
```

```
                Total committed heap usage (bytes)=9179234304
                Peak Map Physical memory (bytes)=398385152
                Peak Map Virtual memory (bytes)=2585186304
                Peak Reduce Physical memory (bytes)=677773312
                Peak Reduce Virtual memory (bytes)=2580742144
        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=155
2024-03-09 16:18:39,524 INFO streaming.StreamJob: Output directory: /assignment1/Output_part2_1_2
hadoop@aminpri-1-2:~$
```

Prithvi Amin (aminpri)
SP24-BL-ENGR-E516-14573

Check output file in browser:





Output:

Top 3 IP addresses from 01:00:00 to 01:59:59:

IP: 66.249.66.91, Count: 13874

IP: 66.249.66.194, Count: 12485

IP: 66.249.66.92, Count: 2924

Similarly, we would be testing for other user inputs

From hour 4 to hour 7:

```
hadoop@aminpri-1-2:~$ hadoop jar /$HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar     -file /home/hadoop
/assignment1/Part_02/mapper.py     -mapper "/usr/bin/python3 /home/hadoop/assignment1/Part_02/mapper.py"     -file /hom
e/hadoop/assignment1/Part_02/reducer.py  -reducer "/usr/bin/python3 /home/hadoop/assignment1/Part_02/reducer.py 3-7"
 -input /assignment1/Input/access.log     -output /assignment1/Output_part2_3_7
2024-03-09 16:24:28,566 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/hadoop/assignment1/Part_02/mapper.py, /home/hadoop/assignment1/Part_02/reducer.py, /tmp/hadoop-un
jar4876869124111305084/] [] /tmp/streamjob2806682920487506583.jar tmpDir=null
2024-03-09 16:24:29,566 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-03-09 16:24:29,960 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-03-09 16:24:30,285 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging
/hadoop/.staging/job_1710011891143_0004
2024-03-09 16:24:31,586 INFO mapred.FileInputFormat: Total input files to process : 1
2024-03-09 16:24:31,611 INFO net.NetworkTopology: Adding a new node: /default-rack/127.0.0.1:9866
2024-03-09 16:24:32,107 INFO mapreduce.JobSubmitter: number of splits:26
2024-03-09 16:24:32,349 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1710011891143_0004
2024-03-09 16:24:32,350 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-03-09 16:24:32,636 INFO conf.Configuration: resource-types.xml not found
2024-03-09 16:24:32,637 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-03-09 16:24:32,725 INFO impl.YarnClientImpl: Submitted application application_1710011891143_0004
2024-03-09 16:24:32,804 INFO mapreduce.Job: The url to track the job: http://aminpri-1-2:8088/proxy/application_1710011
891143_0004/
2024-03-09 16:24:32,806 INFO mapreduce.Job: Running job: job_1710011891143_0004
```

Output:

File contents

Top 3 IP addresses from 03:00:00 to 06:59:59:

IP: 66.249.66.194, Count: 40266

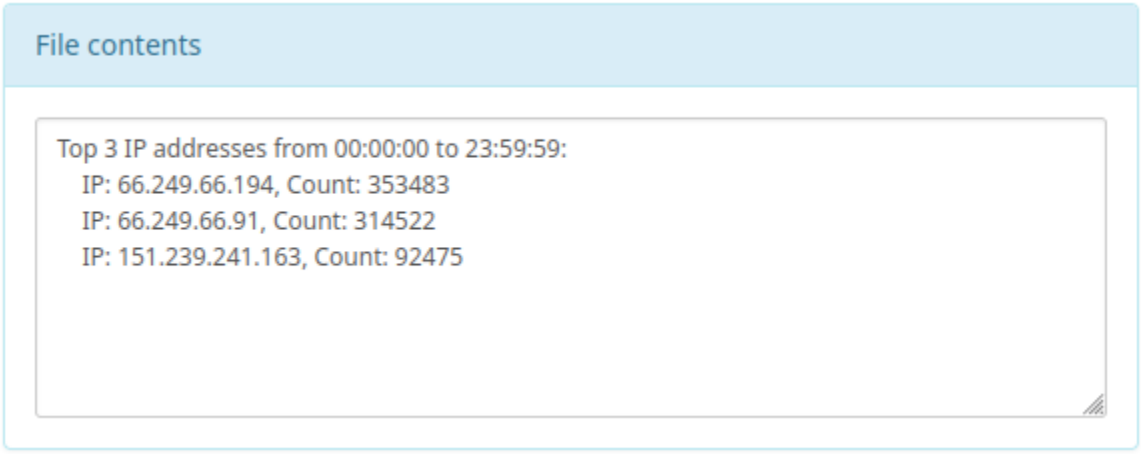
IP: 66.249.66.91, Count: 30488

IP: 91.99.72.15, Count: 6758

From hour 0 to hour 24:

```
hadoop@aminpri-1-2:~$ hadoop jar /$HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar     -file /home/hadoop
/assignment1/Part_02/mapper.py     -mapper "/usr/bin/python3 /home/hadoop/assignment1/Part_02/mapper.py"     -file /hom
e/hadoop/assignment1/Part_02/reducer.py  -reducer "/usr/bin/python3 /home/hadoop/assignment1/Part_02/reducer.py 0-24"
  -input /assignment1/Input/access.log     -output /assignment1/Output_part2_0_24
2024-03-09 16:32:08,555 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/hadoop/assignment1/Part_02/mapper.py, /home/hadoop/assignment1/Part_02/reducer.py, /tmp/hadoop-un
jar3609672301798174956/] [] /tmp/streamjob566048865118688870.jar tmpDir=null
2024-03-09 16:32:09,761 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-03-09 16:32:09,975 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-03-09 16:32:10,424 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging
/hadoop/.staging/job_1710011891143_0005
2024-03-09 16:32:11,065 INFO mapred.FileInputFormat: Total input files to process : 1
2024-03-09 16:32:11,089 INFO net.NetworkTopology: Adding a new node: /default-rack/127.0.0.1:9866
2024-03-09 16:32:11,304 INFO mapreduce.JobSubmitter: number of splits:26
2024-03-09 16:32:11,609 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1710011891143_0005
2024-03-09 16:32:11,609 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-03-09 16:32:11,881 INFO conf.Configuration: resource-types.xml not found
2024-03-09 16:32:11,881 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-03-09 16:32:12,160 INFO impl.YarnClientImpl: Submitted application application_1710011891143_0005
2024-03-09 16:32:12,300 INFO mapreduce.Job: The url to track the job: http://aminpri-1-2:8088/proxy/application_1710011
891143_0005/
2024-03-09 16:32:12,301 INFO mapreduce.Job: Running job: job_1710011891143_0005
```

Output:



Run it along with three other examples, WordCount, Sort, Grep, at the same time, and test fair and capacity schedulers.

Get the location of hadoop mapreduce examples:

```
hadoop@aminpri-1-2:~$ find / -name "hadoop-mapreduce-examples*.jar" -print
find: ‘/root’: Permission denied
find: ‘/lost+found’: Permission denied
/home/hadoop/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.6.jar
/home/hadoop/hadoop/share/hadoop/mapreduce/sources/hadoop-mapreduce-examples-3.3.6-sources.jar
/home/hadoop/hadoop/share/hadoop/mapreduce/sources/hadoop-mapreduce-examples-3.3.6-test-sources.jar
find: ‘/home/aminpri’: Permission denied
find: ‘/proc/tty/driver’: Permission denied
find: ‘/proc/1/task/1/fd’: Permission denied
```

Define environment variable HADOOP_EXAMPLES:

```
hadoop@aminpri-1-2:~$ export HADOOP_EXAMPLES=/home/hadoop/hadoop/share/hadoop/mapreduce
```

Listing available examples:

```
hadoop@aminpri-1-2:~$ yarn jar $HADOOP_EXAMPLES/hadoop-mapreduce-examples-3.3.6.jar
An example program must be given as the first argument.
Valid program names are:
  aggregatewordcount: An Aggregate based map/reduce program that counts the words in the input files.
  aggregatewordhist: An Aggregate based map/reduce program that computes the histogram of the words in the input files.
  bbp: A map/reduce program that uses Bailey-Borwein-Plouffe to compute exact digits of Pi.
  dbcount: An example job that count the pageview counts from a database.
  distbbp: A map/reduce program that uses a BBP-type formula to compute exact bits of Pi.
  grep: A map/reduce program that counts the matches of a regex in the input.
  join: A job that effects a join over sorted, equally partitioned datasets
  multifilewc: A job that counts words from several files.
  pentomino: A map/reduce tile laying program to find solutions to pentomino problems.
  pi: A map/reduce program that estimates Pi using a quasi-Monte Carlo method.
  randomtextwriter: A map/reduce program that writes 10GB of random textual data per node.
  randomwriter: A map/reduce program that writes 10GB of random data per node.
  secondarysort: An example defining a secondary sort to the reduce.
  sort: A map/reduce program that sorts the data written by the random writer.
  sudoku: A sudoku solver.
  teragen: Generate data for the terasort
  terasort: Run the terasort
  teravalidate: Checking results of terasort
  wordcount: A map/reduce program that counts the words in the input files.
  wordmean: A map/reduce program that counts the average length of the words in the input files.
  wordmedian: A map/reduce program that counts the median length of the words in the input files.
  wordstandarddeviation: A map/reduce program that counts the standard deviation of the length of the words in the inpu
t files.
```

Exclusively configuring to use **capacity scheduler** in yarn.xml:

```
hadoop@aminpri-1-2:~$ cat $HADOOP_HOME/etc/hadoop/yarn-site.xml
<?xml version="1.0"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->
<configuration>
<property>
        <name>yarn.nodemanager.aux-services</name>
        <value>mapreduce_shuffle</value>
    </property>

<property>
  <name>yarn.resourcemanager.scheduler.class</name>
  <value>org.apache.hadoop.yarn.server.resourcemanager.scheduler.capacity.CapacityScheduler</value>
</property>

</configuration>
```

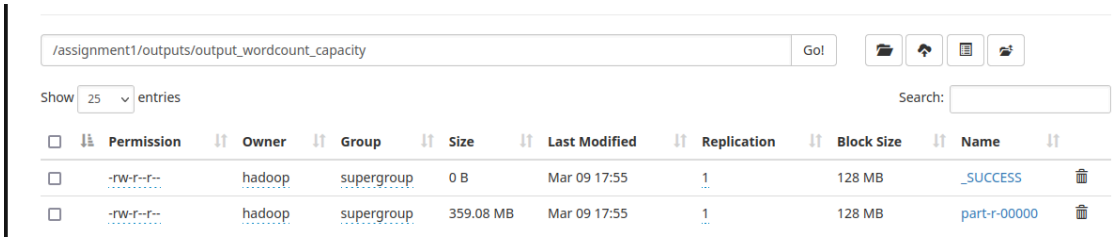
We will now run wordcount, sort and grep using capacity scheduler.

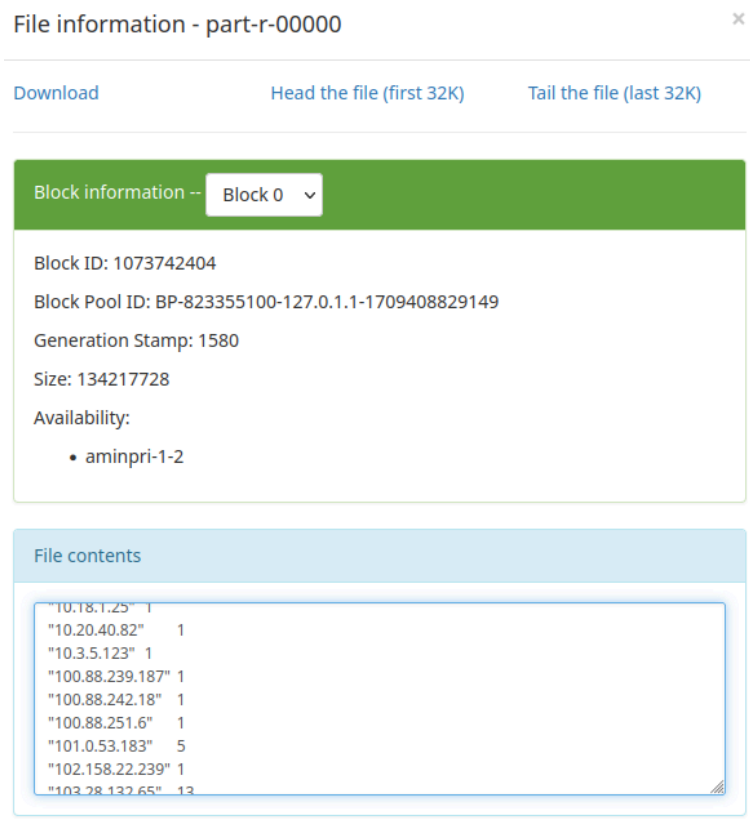
Wordcount using capacity scheduler:

hadoop@aminpri-1-2:~$ yarn jar $HADOOP_EXAMPLES/hadoop-mapreduce-examples-3.3.6.jar wordcount /assignment1/Input/access.log /assignment1/outputs/output_wordcount_capacity

```
hadoop@aminpri-1-2:~$ yarn jar $HADOOP_EXAMPLES/hadoop-mapreduce-examples-3.3.6.jar wordcount /assignment1/Input/access
.log /assignment1/outputs/output_wordcount_capacity
2024-03-09 17:50:55,094 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-03-09 17:50:55,692 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging
/hadoop/.staging/job_1710011891143_0007
2024-03-09 17:50:56,196 INFO input.FileInputFormat: Total input files to process : 1
2024-03-09 17:50:56,423 INFO mapreduce.JobSubmitter: number of splits:26
2024-03-09 17:50:56,665 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1710011891143_0007
2024-03-09 17:50:56,665 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-03-09 17:50:56,964 INFO conf.Configuration: resource-types.xml not found
2024-03-09 17:50:56,964 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-03-09 17:50:57,087 INFO impl.YarnClientImpl: Submitted application application_1710011891143_0007
2024-03-09 17:50:57,155 INFO mapreduce.Job: The url to track the job: http://aminpri-1-2:8088/proxy/application_1710011
891143_0007/
2024-03-09 17:50:57,156 INFO mapreduce.Job: Running job: job_1710011891143_0007
2024-03-09 17:51:09,518 INFO mapreduce.Job: Job job_1710011891143_0007 running in uber mode : false
2024-03-09 17:51:09,519 INFO mapreduce.Job:  map 0% reduce 0%
```

Output:



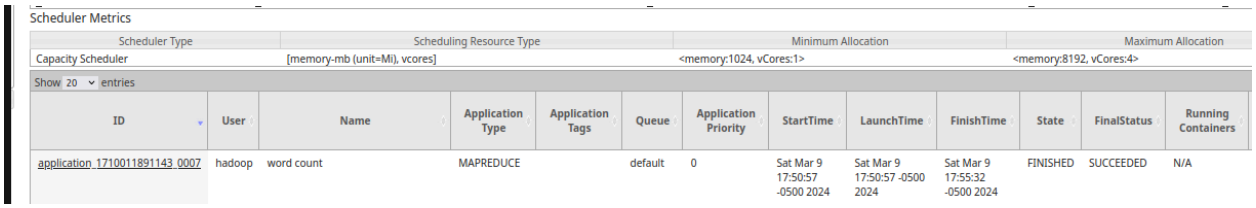


Prithvi Amin (aminpri)
SP24-BL-ENGR-E516-14573

Head of the output file:

" 1
"" 33
"&\xBEi\x86p\xBAOt}U\x95\xB3\x5C\x1F\x02\x1D\x84\xD7\xAA{3]\x1B\x9BK\xE2\xEF\xF79\xBB\xFD\x05g\xFB\x1B\xE5\xFD1\xCEJ\xFB\xD0\xBE\xF8/\xBD_y\xE0WX\xF7,]\xDC\xBD\xFE\xE5\xD5\x7Fr\xA5\x9B\xFB\x97H[v\x7F\xF3\xE2\xEELa\xCA:\x10\xEE\xC5^K\xE0\xBF\x09\xE0Uy\x1F\x81[\x5C\x8C\xF8\xA3<\x90\xF8\xAB\x1F\xD8\xDF$\xE0p\x96\xFF\x00\xDC\x11\xE4v3\xD2\x98\xA8" 1
"(null)" 1
"+https://www.adbeat.com/policy 4
"-" 12001922
"/android-app://com.google.android.gm" 2
"/android-app://com.google.android.googlequicksearchbox" 5
"/android-app://com.google.android.googlequicksearchbox/https/www.google.com" 3
"/android-app://org.telegram.messenger" 1
"10.115.0.93" 9
"10.115.0.97" 3
"10.115.1.205" 6
"10.115.21.73" 3
"10.115.21.75" 3
"10.115.21.77" 6
"10.16.209.128" 1
"10.16.209.241" 1
"10.18.1.25" 1
"10.20.40.82" 1
"10.3.5.123" 1
"100.88.239.187" 1

Metrics:



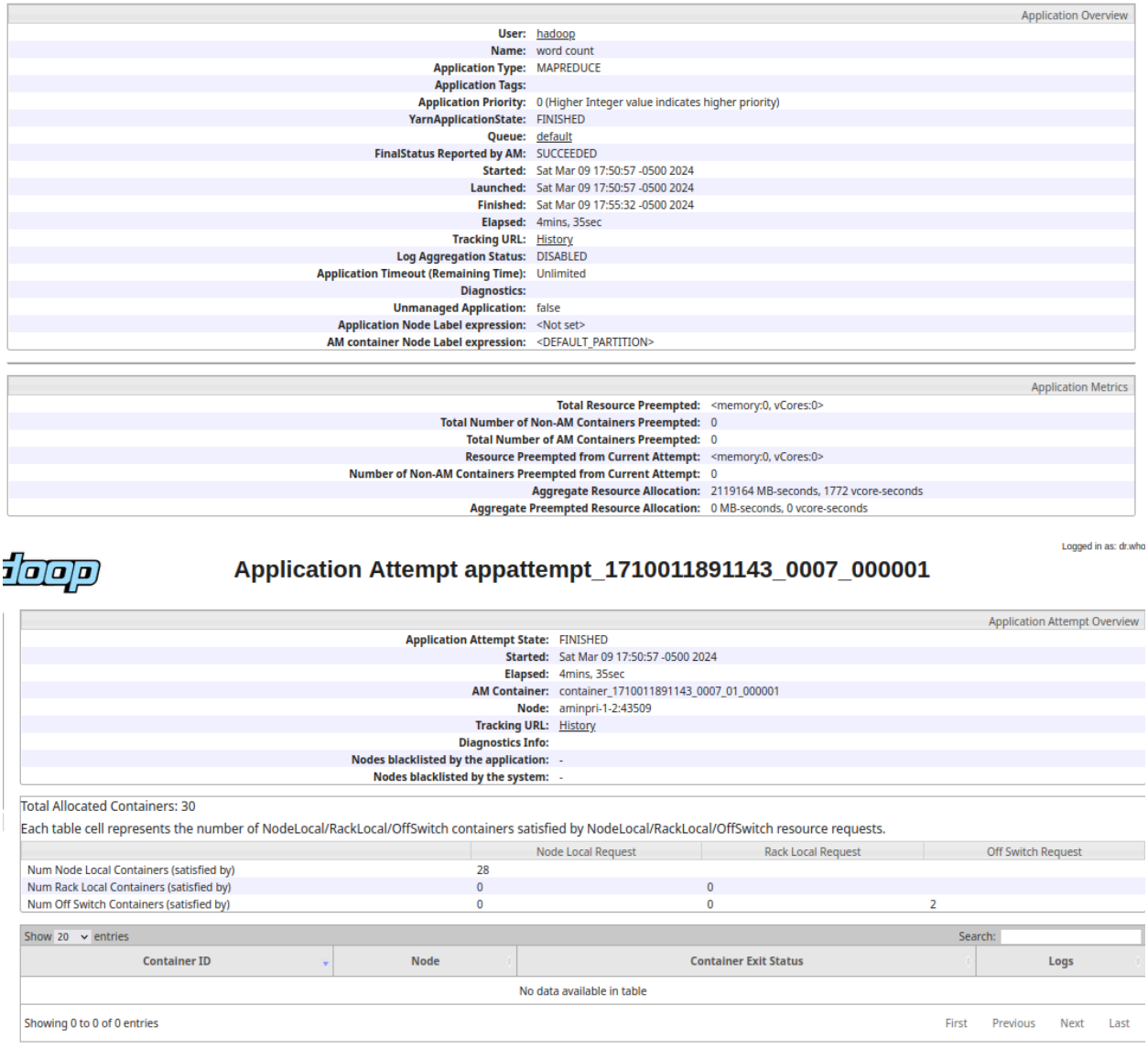
Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers
application_1710011891143_0007	hadoop	word count	MAPREDUCE		default	0	Sat Mar 9 17:50:57 -0500 2024	Sat Mar 9 17:50:57 -0500 2024	Sat Mar 9 17:55:32 -0500 2024	FINISHED	SUCCEEDED	N/A

Prithvi Amin (aminpri)
SP24-BL-ENGR-E516-14573



Application Overview

User:	hadoop
Name:	word count
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sat Mar 09 17:50:57 -0500 2024
Launched:	Sat Mar 09 17:50:57 -0500 2024
Finished:	Sat Mar 09 17:55:32 -0500 2024
Elapsed:	4mins, 35sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

Application Metrics

Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt:	0
Aggregate Resource Allocation:	2119164 MB-seconds, 1772 vcore-seconds
Aggregate Preempted Resource Allocation:	0 MB-seconds, 0 vcore-seconds

Logged in as: dr.who

Application Attempt appattempt_1710011891143_0007_000001

Application Attempt Overview

Application Attempt State:	FINISHED
Started:	Sat Mar 09 17:50:57 -0500 2024
Elapsed:	4mins, 35sec
AM Container:	container_1710011891143_0007_01_000001
Node:	aminpri-1-2:43509
Tracking URL:	History
Diagnostics Info:	
Nodes blacklisted by the application:	-
Nodes blacklisted by the system:	-

Total Allocated Containers: 30

Each table cell represents the number of NodeLocal/RackLocal/OffSwitch containers satisfied by NodeLocal/RackLocal/OffSwitch resource requests.

	Node Local Request	Rack Local Request	Off Switch Request
Num Node Local Containers (satisfied by)	28		
Num Rack Local Containers (satisfied by)	0	0	
Num Off Switch Containers (satisfied by)	0	0	2

Show 20 entries Search:

Container ID	Node	Container Exit Status	Logs
No data available in table			

Showing 0 to 0 of 0 entries First Previous Next Last

Capacity Scheduler Metrics:

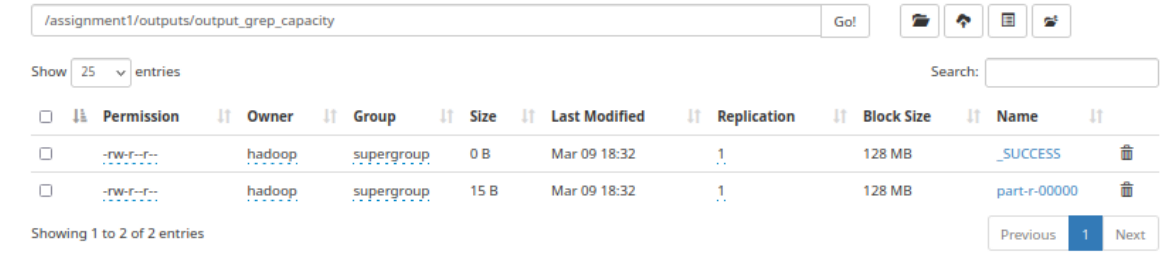
Application ID: application_1710011891143_0007
Elapsed Time: 4 minutes, 35 seconds
Total Allocated Containers: 30
Aggregate Resource Allocation: 2,119,164 MB-seconds, 1,772 vcore-seconds
Start Time: Sat Mar 9 17:50:57 -0500 2024
Finish Time: Sat Mar 9 17:55:32 -0500 2024
Node Local Containers: 28
Rack Local Containers: 0
Off-Switch Containers: 2

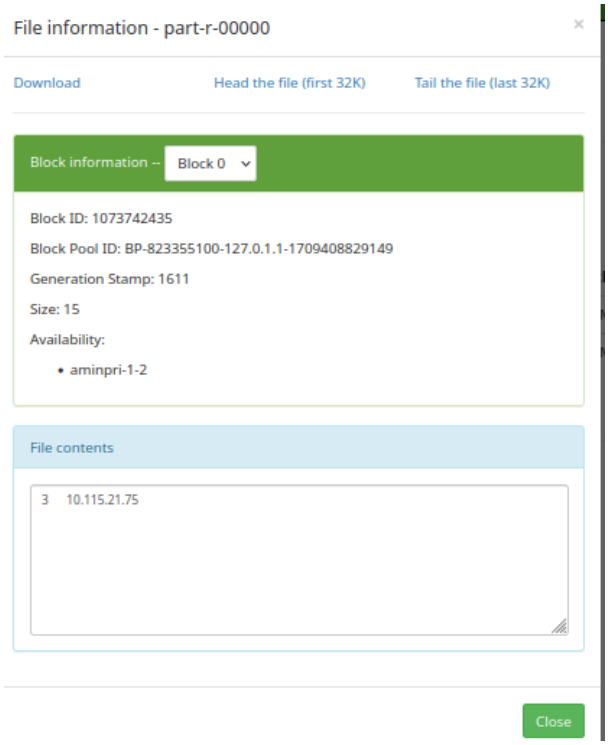
Grep using capacity scheduler:

hadoop@aminpri-1-2:~$ yarn jar $HADOOP_EXAMPLES/hadoop-mapreduce-examples-3.3.6.jar grep /assignment1/Input/access.log /assignment1/outputs/output_grep_capacity "10.115.21.75"

```
hadoop@aminpri-1-2:~$ yarn jar $HADOOP_EXAMPLES/hadoop-mapreduce-examples-3.3.6.jar grep /assignment1/Input/access.log
/assignment1/outputs/output_grep_capacity "10.115.21.75"
2024-03-09 18:29:59,236 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-03-09 18:29:59,775 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging
/hadoop/.staging/job_1710011891143_0009
2024-03-09 18:30:00,207 INFO input.FileInputFormat: Total input files to process : 1
2024-03-09 18:30:00,400 INFO mapreduce.JobSubmitter: number of splits:26
2024-03-09 18:30:00,591 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1710011891143_0009
2024-03-09 18:30:00,591 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-03-09 18:30:00,854 INFO conf.Configuration: resource-types.xml not found
2024-03-09 18:30:00,855 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-03-09 18:30:01,017 INFO impl.YarnClientImpl: Submitted application application_1710011891143_0009
2024-03-09 18:30:01,151 INFO mapreduce.Job: The url to track the job: http://aminpri-1-2:8088/proxy/application_1710011
891143_0009/
2024-03-09 18:30:01,152 INFO mapreduce.Job: Running job: job_1710011891143_0009
2024-03-09 18:30:09,416 INFO mapreduce.Job: Job job_1710011891143_0009 running in uber mode : false
2024-03-09 18:30:09,417 INFO mapreduce.Job:  map 0% reduce 0%
2024-03-09 18:30:29,968 INFO mapreduce.Job:  map 12% reduce 0%
2024-03-09 18:30:30,994 INFO mapreduce.Job:  map 23% reduce 0%
2024-03-09 18:30:41,128 INFO mapreduce.Job:  map 27% reduce 0%
2024-03-09 18:30:42,369 INFO mapreduce.Job:  map 38% reduce 0%
```

Output:





Prithvi Amin (aminpri)
SP24-BL-ENGR-E516-14573

Metrics:

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers
application_1710011891143_0010	hadoop	grep-sort	MAPREDUCE		default	0	Sat Mar 9 18:32:16 -0500 2024	Sat Mar 9 18:32:21 -0500 2024	Sat Mar 9 18:32:35 -0500 2024	FINISHED	SUCCEEDED	N/A
application_1710011891143_0009	hadoop	grep-search	MAPREDUCE		default	0	Sat Mar 9 18:30:00 -0500 2024	Sat Mar 9 18:30:01 -0500 2024	Sat Mar 9 18:32:14 -0500 2024	FINISHED	SUCCEEDED	N/A

Grep-search: application_1710011891143_0009

Logged in as: dr.who

Application application_1710011891143_0009

Application Overview

User:	hadoop
Name:	grep-search
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sat Mar 09 18:30:00 -0500 2024
Launched:	Sat Mar 09 18:30:01 -0500 2024
Finished:	Sat Mar 09 18:32:14 -0500 2024
Elapsed:	2mins, 13sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

Application Metrics

Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt:	0
Aggregate Resource Allocation:	1026215 MB-seconds, 847 vcore-seconds
Aggregate Preempted Resource Allocation:	0 MB-seconds, 0 vcore-seconds

Show 20 entries Search:

Attempt ID	Started	Node	Logs	Nodes blacklisted by the app	Nodes blacklisted by the system
appattempt_1710011891143_0009_000001	Sat Mar 9 18:30:00 -0500 2024	http://aminpri-1-2:8042	Logs	0	0

Logged in as: dr.who

Application Attempt appattempt_1710011891143_0009_000001

Application Attempt Overview

Application Attempt State:	FINISHED
Started:	Sat Mar 09 18:30:00 -0500 2024
Elapsed:	2mins, 13sec
AM Container:	container_1710011891143_0009_01_000001
Node:	aminpri-1-2:44741
Tracking URL:	History
Diagnostics Info:	
Nodes blacklisted by the application:	-
Nodes blacklisted by the system:	-

Total Allocated Containers: 30

Each table cell represents the number of NodeLocal/RackLocal/OffSwitch containers satisfied by NodeLocal/RackLocal/OffSwitch resource requests.

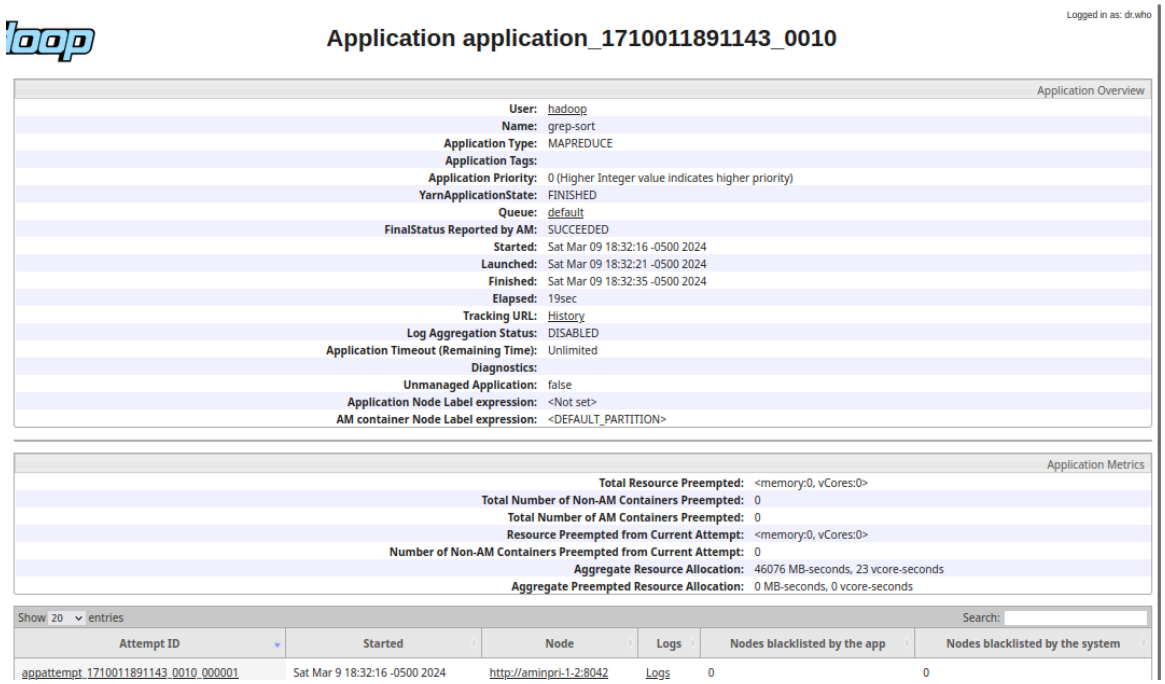
	Node Local Request	Rack Local Request	Off Switch Request
Num Node Local Containers (satisfied by)	28		
Num Rack Local Containers (satisfied by)	0	0	
Num Off Switch Containers (satisfied by)	0	0	2

Show 20 entries Search:

Container ID	Node	Container Exit Status	Logs
No data available in table			

Showing 0 to 0 of 0 entries First Previous Next Last

grep-sort: application_1710011891143_0010



Logged in as: dr.who

Application application_1710011891143_0010

Application Overview

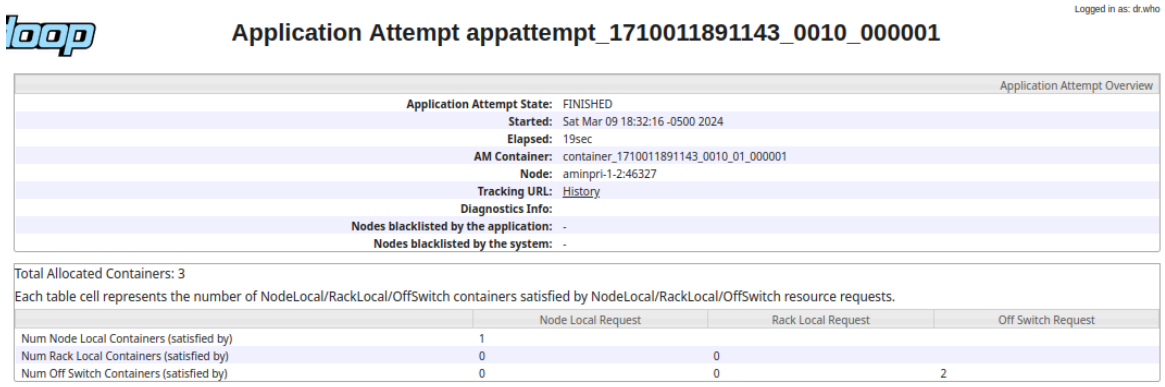
User:	hadoop
Name:	grep-sort
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sat Mar 09 18:32:16 -0500 2024
Launched:	Sat Mar 09 18:32:21 -0500 2024
Finished:	Sat Mar 09 18:32:35 -0500 2024
Elapsed:	19sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

Application Metrics

Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt:	0
Aggregate Resource Allocation:	46076 MB-seconds, 23 vcore-seconds
Aggregate Preempted Resource Allocation:	0 MB-seconds, 0 vcore-seconds

Show 20 entries Search:

Attempt ID	Started	Node	Logs	Nodes blacklisted by the app	Nodes blacklisted by the system
appattempt_1710011891143_0010_000001	Sat Mar 9 18:32:16 -0500 2024	http://aminpri-1-2:8042	Logs	0	0



Logged in as: dr.who

Application Attempt appattempt_1710011891143_0010_000001

Application Attempt Overview

Application Attempt State:	FINISHED
Started:	Sat Mar 09 18:32:16 -0500 2024
Elapsed:	19sec
AM Container:	container_1710011891143_0010_01_000001
Node:	aminpri-1-2:46327
Tracking URL:	History
Diagnostics Info:	
Nodes blacklisted by the application:	-
Nodes blacklisted by the system:	-

Total Allocated Containers: 3

Each table cell represents the number of NodeLocal/RackLocal/OffSwitch containers satisfied by NodeLocal/RackLocal/OffSwitch resource requests.

	Node Local Request	Rack Local Request	Off Switch Request
Num Node Local Containers (satisfied by)	1		
Num Rack Local Containers (satisfied by)	0	0	
Num Off Switch Containers (satisfied by)	0	0	2

Capacity Scheduler Metrics for grep-search:
Application ID: application_1710011891143_0009
Elapsed Time: 2 minutes, 13 seconds
Total Allocated Containers: 30
Aggregate Resource Allocation: 1,026,215 MB-seconds, 847 vcore-seconds

Capacity Scheduler Metrics for grep-sort:
Application ID: application_1710011891143_0010
Elapsed Time: 19 seconds
Total Allocated Containers: 3
Aggregate Resource Allocation: 406,676 MB-seconds, 284 vcore-seconds

Sort using capacity scheduler:
Hadoop's sort job requires a Sequence file as input, and access.log is not a Sequence file. Therefore, we will need to write our own code for the mapper and reducer for sorting.

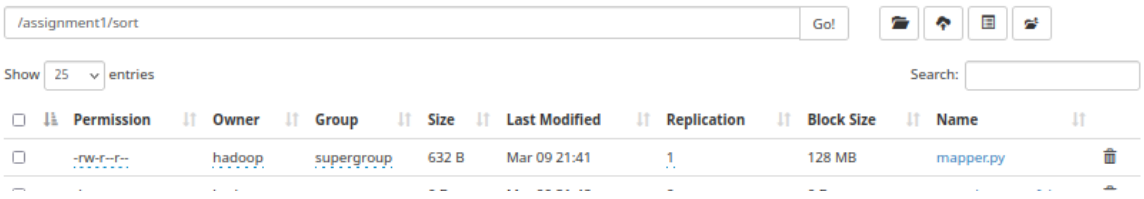
Create the file:

```
hadoop@aminpri-1-2:~/assignment1/sort$ ls
mapper.py
hadoop@aminpri-1-2:~/assignment1/sort$ cat mapper.py
#!/usr/bin/env python3

import sys

# Function to parse the log entry and extract the IP address
def get_ip_address(log_entry):
    parts = log_entry.split(' ')
    return parts[0]

# Function to read from stdin, sort log entries based on IP addresses, and write to stdout
def sort_logs():
    # Read input data from stdin
    lines = sys.stdin.readlines()
    # Sort the log entries based on IP addresses
    sorted_lines = sorted(lines, key=get_ip_address)
    # Write sorted log entries to stdout
    for line in sorted_lines:
        print(line, end='')

# Call the sort_logs function
if __name__ == "__main__":
    sort_logs()
```

Mount to hdfs:

```
hadoop@aminpri-1-2:~/assignment1/sort$ cd
hadoop@aminpri-1-2:~$ hdfs dfs -copyFromLocal "assignment1/sort" /assignment1
```

Check in browser:

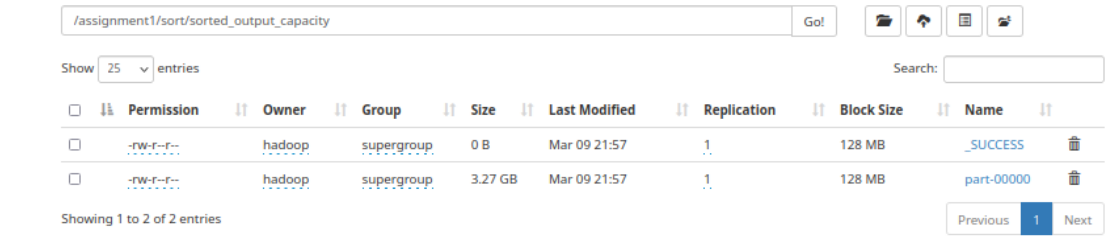


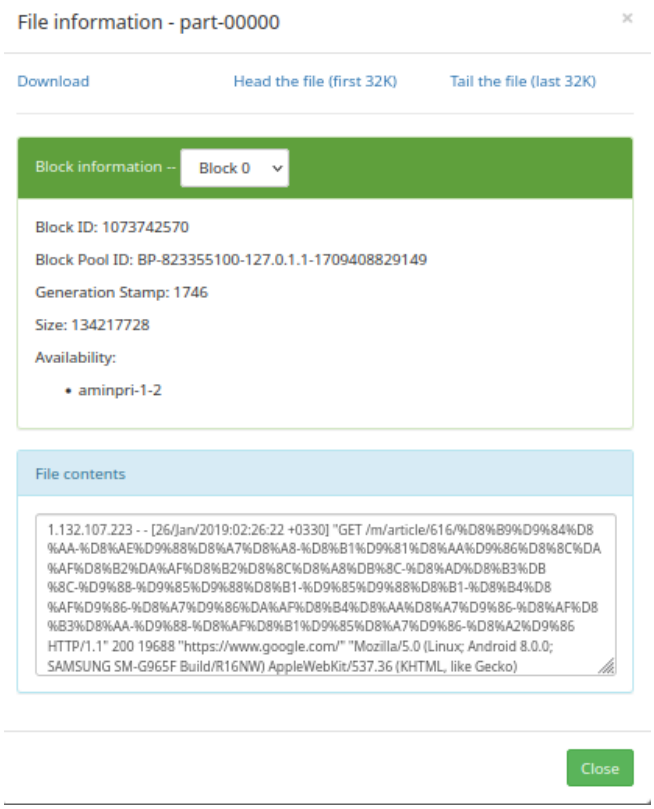
Execute the below command:

hadoop@aminpri-1-2:~$ hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar -files /home/hadoop/assignment1/sort/mapper.py -input /assignment1/Input/access.log -output /assignment1/sort/sorted_output_capacity -mapper "python3 /home/hadoop/assignment1/sort/mapper.py" -reducer "cat"

```
hadoop@aminpri-1-2:~$ hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar     -files /home/hadoop
/assignment1/sort/mapper.py     -input /assignment1/Input/access.log     -output /assignment1/sort/sorted_output_capaci
ty     -mapper "python3 /home/hadoop/assignment1/sort/mapper.py"     -reducer "cat"
packageJobJar: [/tmp/hadoop-unjar6275432407454652467/] [] /tmp/streamjob7648933284233449466.jar tmpDir=null
2024-03-09 21:52:35,170 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-03-09 21:52:35,790 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-03-09 21:52:37,212 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging
/hadoop/.staging/job_1710011891143_0020
2024-03-09 21:52:38,992 INFO mapred.FileInputFormat: Total input files to process : 1
2024-03-09 21:52:39,247 INFO net.NetworkTopology: Adding a new node: /default-rack/127.0.0.1:9866
2024-03-09 21:52:39,851 INFO mapreduce.JobSubmitter: number of splits:26
2024-03-09 21:52:40,568 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1710011891143_0020
2024-03-09 21:52:40,578 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-03-09 21:52:41,664 INFO conf.Configuration: resource-types.xml not found
2024-03-09 21:52:41,664 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-03-09 21:52:41,762 INFO impl.YarnClientImpl: Submitted application application_1710011891143_0020
2024-03-09 21:52:41,818 INFO mapreduce.Job: The url to track the job: http://aminpri-1-2:8088/proxy/application_1710011
891143_0020/
2024-03-09 21:52:41,820 INFO mapreduce.Job: Running job: job_1710011891143_0020
2024-03-09 21:52:59,643 INFO mapreduce.Job: Job job_1710011891143_0020 running in uber mode : false
2024-03-09 21:52:59,644 INFO mapreduce.Job:  map 0% reduce 0%
2024-03-09 21:53:27,525 INFO mapreduce.Job:  map 1% reduce 0%
2024-03-09 21:53:28,574 INFO mapreduce.Job:  map 5% reduce 0%
```

Output:





Metrics:

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers
application_1710011891143_0020	hadoop	streamjob7648933284233449466.jar	MAPREDUCE		default	0	Sat Mar 9 21:52:41 -0500 2024	Sat Mar 9 21:52:42 -0500 2024	Sat Mar 9 21:57:36 -0500 2024	FINISHED	SUCCEEDED	N/A

Logged in as: dr.who

Application application_1710011891143_0020

Application Overview

User:	hadoop
Name:	streamjob7648933284233449466.jar
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sat Mar 09 21:52:41 -0500 2024
Launched:	Sat Mar 09 21:52:42 -0500 2024
Finished:	Sat Mar 09 21:57:36 -0500 2024
Elapsed:	4mins, 54sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

Application Metrics

Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt:	0
Aggregate Resource Allocation:	1944639 MB-seconds, 1584 vcore-seconds
Aggregate Preempted Resource Allocation:	0 MB-seconds, 0 vcore-seconds

Show 20 entries Search:

Attempt ID	Started	Node	Logs	Nodes blacklisted by the app	Nodes blacklisted by the system
appattempt_1710011891143_0020_000001	Sat Mar 9 21:52:41 -0500 2024	http://aminpri-1-2:8042	Logs	0	0

Logged in as: dr.who

Application Attempt appattempt_1710011891143_0020_000001

Application Attempt Overview

Application Attempt State:	FINISHED
Started:	Sat Mar 09 21:52:41 -0500 2024
Elapsed:	4mins, 54sec
AM Container:	container_1710011891143_0020_01_000001
Node:	aminpri-1-2:36347
Tracking URL:	History
Diagnostics Info:	
Nodes blacklisted by the application:	-
Nodes blacklisted by the system:	-

Total Allocated Containers: 29

Each table cell represents the number of NodeLocal/RackLocal/OffSwitch containers satisfied by NodeLocal/RackLocal/OffSwitch resource requests.

	Node Local Request	Rack Local Request	Off Switch Request
Num Node Local Containers (satisfied by)	26		
Num Rack Local Containers (satisfied by)	0	1	
Num Off Switch Containers (satisfied by)	0	0	2

Capacity Scheduler Metrics:

Application ID: application_1710011891143_0020
Elapsed Time: 4 minutes, 54 seconds
Total Allocated Containers: 29
Aggregate Resource Allocation: 1,944,639 MB-seconds, 1,584 vcore-seconds

Exclusively configuring to use **fair scheduler** in yarn.xml:

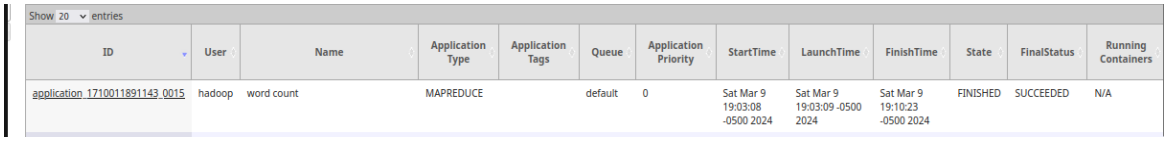
```
hadoop@aminpri-1-2:~$ cat $HADOOP_HOME/etc/hadoop/yarn-site.xml
<?xml version="1.0"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->
<configuration>
<property>
        <name>yarn.nodemanager.aux-services</name>
        <value>mapreduce_shuffle</value>
    </property>

<property>
  <name>yarn.resourcemanager.scheduler.class</name>
  <value>org.apache.hadoop.yarn.server.resourcemanager.scheduler.fair.FairScheduler</value>
</property>

</configuration>
```

We will now run wordcount, sort and grep using a fair scheduler.

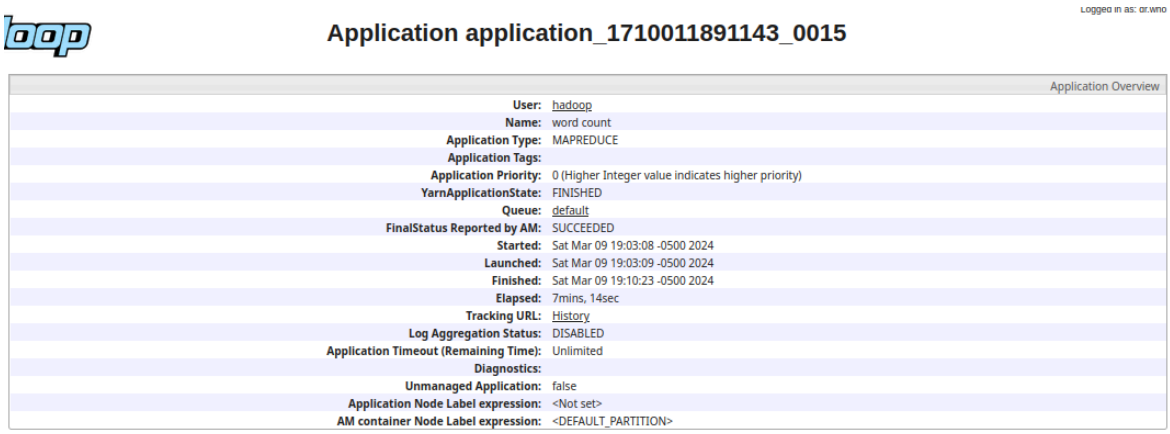
Wordcount using fair scheduler:
hadoop@aminpri-1-2:~$ yarn jar $HADOOP_EXAMPLES/hadoop-mapreduce-examples-3.3.6.jar wordcount /assignment1/Input/access.log /assignment1/outputs/output_wordcount_fair

```
hadoop@aminpri-1-2:~$ yarn jar $HADOOP_EXAMPLES/hadoop-mapreduce-examples-3.3.6.jar wordcount /assignment1/Input/access
.log /assignment1/outputs/output_wordcount_fair
2024-03-09 19:03:01,156 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-03-09 19:03:02,783 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging
/hadoop/.staging/job_1710011891143_0015
2024-03-09 19:03:04,062 INFO input.FileInputFormat: Total input files to process : 1
2024-03-09 19:03:05,309 INFO mapreduce.JobSubmitter: number of splits:26
2024-03-09 19:03:06,482 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1710011891143_0015
2024-03-09 19:03:06,482 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-03-09 19:03:08,306 INFO conf.Configuration: resource-types.xml not found
2024-03-09 19:03:08,323 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-03-09 19:03:08,782 INFO impl.YarnClientImpl: Submitted application application_1710011891143_0015
2024-03-09 19:03:08,942 INFO mapreduce.Job: The url to track the job: http://aminpri-1-2:8088/proxy/application_1710011
891143_0015/
2024-03-09 19:03:08,948 INFO mapreduce.Job: Running job: job_1710011891143_0015
2024-03-09 19:03:29,300 INFO mapreduce.Job: Job job_1710011891143_0015 running in uber mode : false
2024-03-09 19:03:29,302 INFO mapreduce.Job:  map 0% reduce 0%
2024-03-09 19:04:07,709 INFO mapreduce.Job:  map 1% reduce 0%
2024-03-09 19:04:08,887 INFO mapreduce.Job:  map 2% reduce 0%
2024-03-09 19:04:10,424 INFO mapreduce.Job:  map 4% reduce 0%
2024-03-09 19:04:15,997 INFO mapreduce.Job:  map 5% reduce 0%
2024-03-09 19:04:23,563 INFO mapreduce.Job:  map 6% reduce 0%
```

Metrics:



ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers
application_1710011891143_0015	hadoop	word count	MAPREDUCE		default	0	Sat Mar 9 19:03:08 -0500 2024	Sat Mar 9 19:03:09 -0500 2024	Sat Mar 9 19:10:23 -0500 2024	FINISHED	SUCCEEDED	N/A

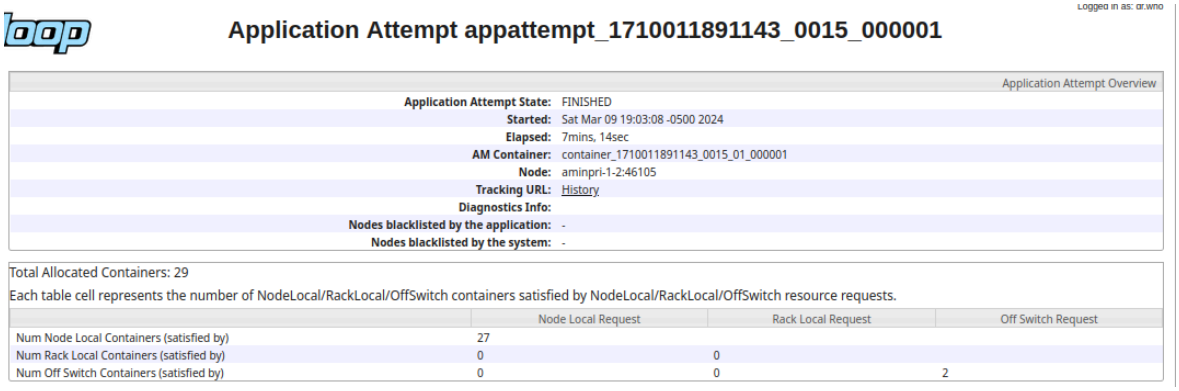


Application application_1710011891143_0015

Application Overview	
User:	hadoop
Name:	word count
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sat Mar 09 19:03:08 -0500 2024
Launched:	Sat Mar 09 19:03:09 -0500 2024
Finished:	Sat Mar 09 19:10:23 -0500 2024
Elapsed:	7mins, 14sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

Application Metrics	
Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt:	0
Aggregate Resource Allocation:	3374976 MB-seconds, 2840 vcore-seconds
Aggregate Preempted Resource Allocation:	0 MB-seconds, 0 vcore-seconds

Attempt ID	Started	Node	Logs	Nodes blacklisted by the app	Nodes blacklisted by the system
appattempt_1710011891143_0015_000001	Sat Mar 9 19:03:08 -0500 2024	http://aminpri-1-2:8042	Logs	0	0



Application Attempt appattempt_1710011891143_0015_000001

Application Attempt Overview	
Application Attempt State:	FINISHED
Started:	Sat Mar 09 19:03:08 -0500 2024
Elapsed:	7mins, 14sec
AM Container:	container_1710011891143_0015_01_000001
Node:	aminpri-1-2:46105
Tracking URL:	History
Diagnostics Info:	
Nodes blacklisted by the application:	-
Nodes blacklisted by the system:	-

Total Allocated Containers: 29

Each table cell represents the number of NodeLocal/RackLocal/OffSwitch containers satisfied by NodeLocal/RackLocal/OffSwitch resource requests.

	Node Local Request	Rack Local Request	Off Switch Request
Num Node Local Containers (satisfied by)	27		
Num Rack Local Containers (satisfied by)	0	0	
Num Off Switch Containers (satisfied by)	0	0	2

Fair Scheduler Metrics:

Application ID: application_1710011891143_0015
Elapsed Time: 7 minutes, 14 seconds
Total Allocated Containers: 29
Aggregate Resource Allocation: 3,374,976 MB-seconds, 2,840 vcore-seconds
Start Time: Sat Mar 9 19:03:08 -0500 2024
Finish Time: Sat Mar 9 19:10:23 -0500 2024
Node Local Containers: 27
Rack Local Containers: 0
Off-Switch Containers: 2

Grep using fair scheduler:

hadoop@aminpri-1-2:~$ yarn jar $HADOOP_EXAMPLES/hadoop-mapreduce-examples-3.3.6.jar grep /assignment1/Input/access.log /assignment1/outputs/output_grep_fair "10.115.21.75"

```
hadoop@aminpri-1-2:~$ yarn jar $HADOOP_EXAMPLES/hadoop-mapreduce-examples-3.3.6.jar grep /assignment1/Input/access.log
/assignment1/outputs/output_grep_fair "10.115.21.75"
2024-03-09 19:16:41,125 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-03-09 19:16:45,839 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging
/hadoop/.staging/job_1710011891143_0016
2024-03-09 19:16:48,797 INFO input.FileInputFormat: Total input files to process : 1
2024-03-09 19:16:50,091 INFO mapreduce.JobSubmitter: number of splits:26
2024-03-09 19:16:52,767 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1710011891143_0016
2024-03-09 19:16:52,767 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-03-09 19:16:54,855 INFO conf.Configuration: resource-types.xml not found
2024-03-09 19:16:54,856 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-03-09 19:16:56,774 INFO impl.YarnClientImpl: Submitted application application_1710011891143_0016
2024-03-09 19:16:57,666 INFO mapreduce.Job: The url to track the job: http://aminpri-1-2:8088/proxy/application_1710011
891143_0016/
2024-03-09 19:16:57,667 INFO mapreduce.Job: Running job: job_1710011891143_0016
2024-03-09 19:18:14,694 INFO mapreduce.Job: Job job_1710011891143_0016 running in uber mode : false
2024-03-09 19:18:14,835 INFO mapreduce.Job:  map 0% reduce 0%
2024-03-09 19:18:39,280 INFO mapreduce.Job:  map 4% reduce 0%
2024-03-09 19:18:40,302 INFO mapreduce.Job:  map 20% reduce 0%
2024-03-09 19:18:41,342 INFO mapreduce.Job:  map 23% reduce 0%
2024-03-09 19:19:00,421 INFO mapreduce.Job:  map 27% reduce 0%
2024-03-09 19:19:03,019 INFO mapreduce.Job:  map 29% reduce 0%
2024-03-09 19:19:04,326 INFO mapreduce.Job:  map 42% reduce 0%
```

Metrics:

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers
application_1710011891143_0017	hadoop	grep-sort	MAPREDUCE		default	0	Sat Mar 9 19:20:02 -0500 2024	Sat Mar 9 19:20:06 -0500 2024	Sat Mar 9 19:20:26 -0500 2024	FINISHED	SUCCEEDED	N/A
application_1710011891143_0016	hadoop	grep-search	MAPREDUCE		default	0	Sat Mar 9 19:16:55 -0500 2024	Sat Mar 9 19:16:57 -0500 2024	Sat Mar 9 19:19:59 -0500 2024	FINISHED	SUCCEEDED	N/A

Grep-search: application_1710011891143_0016

Logged in as: dr.who

Application application_1710011891143_0016

Application Overview

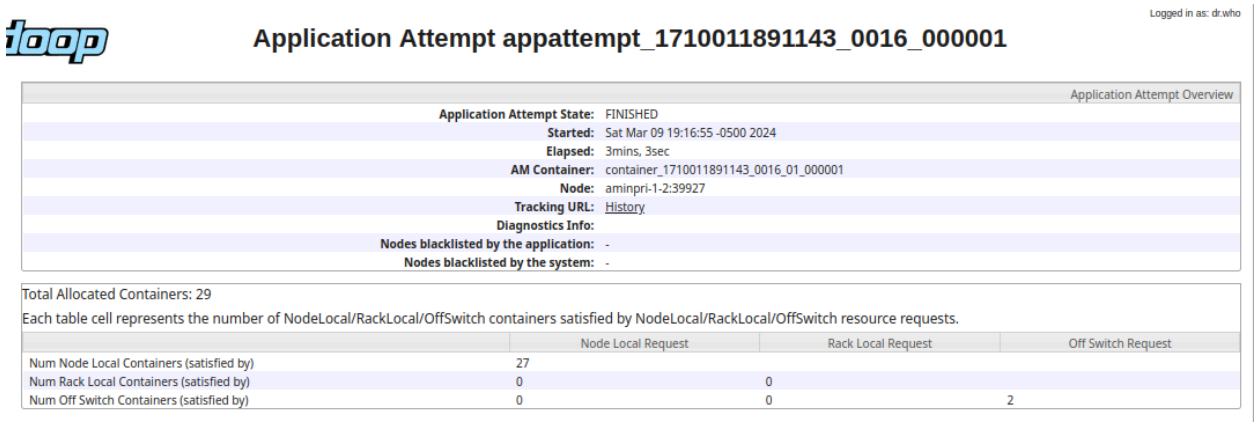
User:	hadoop
Name:	grep-search
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sat Mar 09 19:16:55 -0500 2024
Launched:	Sat Mar 09 19:16:57 -0500 2024
Finished:	Sat Mar 09 19:19:59 -0500 2024
Elapsed:	3mins, 3sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

Application Metrics

Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt:	0
Aggregate Resource Allocation:	997787 MB-seconds, 772 vcore-seconds
Aggregate Preempted Resource Allocation:	0 MB-seconds, 0 vcore-seconds

Show 20 entries Search:

Attempt ID	Started	Node	Logs	Nodes blacklisted by the app	Nodes blacklisted by the system
appattempt_1710011891143_0016_000001	Sat Mar 9 19:16:55 -0500 2024	http://aminpri-1-2:8042	Logs	0	0



Logged in as: dr.who

Application Attempt appattempt_1710011891143_0016_000001

Application Attempt Overview

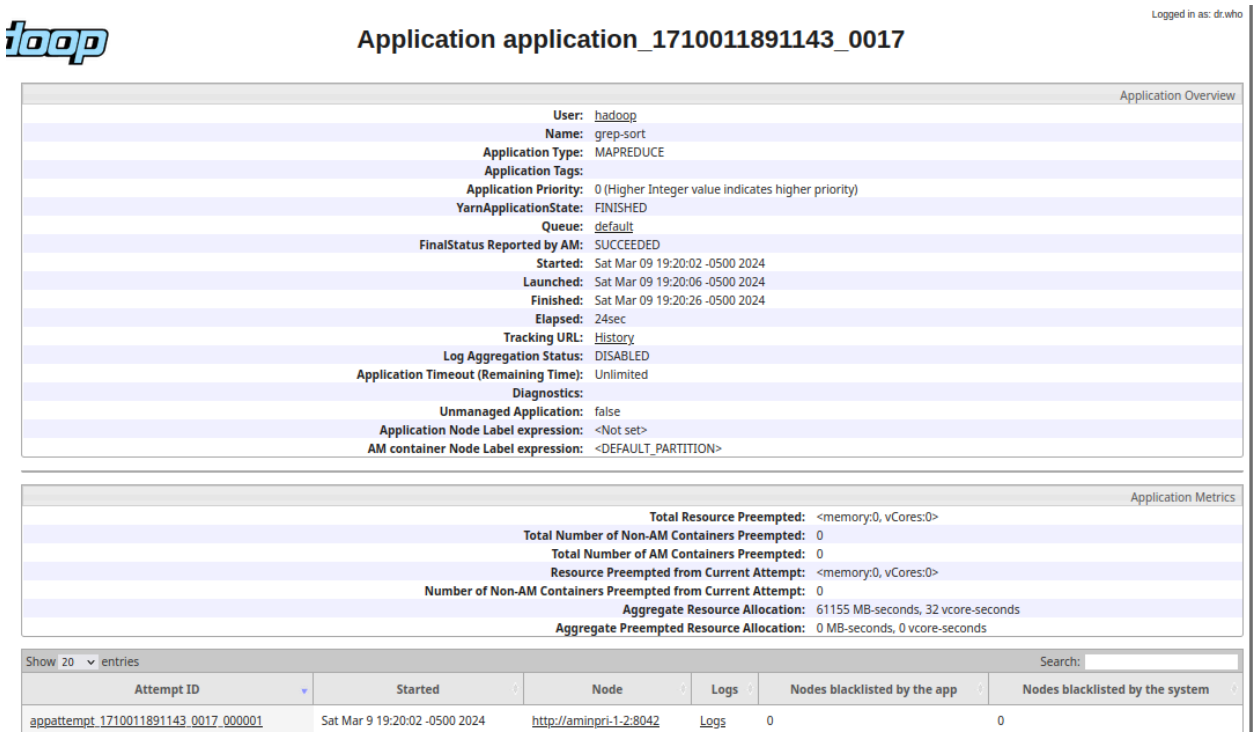
Application Attempt State:	FINISHED
Started:	Sat Mar 09 19:16:55 -0500 2024
Elapsed:	3mins, 3sec
AM Container:	container_1710011891143_0016_01_000001
Node:	aminpri-1-2:39927
Tracking URL:	History
Diagnostics Info:	
Nodes blacklisted by the application:	-
Nodes blacklisted by the system:	-

Total Allocated Containers: 29

Each table cell represents the number of NodeLocal/RackLocal/OffSwitch containers satisfied by NodeLocal/RackLocal/OffSwitch resource requests.

	Node Local Request	Rack Local Request	Off Switch Request
Num Node Local Containers (satisfied by)	27		
Num Rack Local Containers (satisfied by)	0	0	
Num Off Switch Containers (satisfied by)	0	0	2

Grep-sort: application_1710011891143_0017



Logged in as: dr.who

Application application_1710011891143_0017

Application Overview

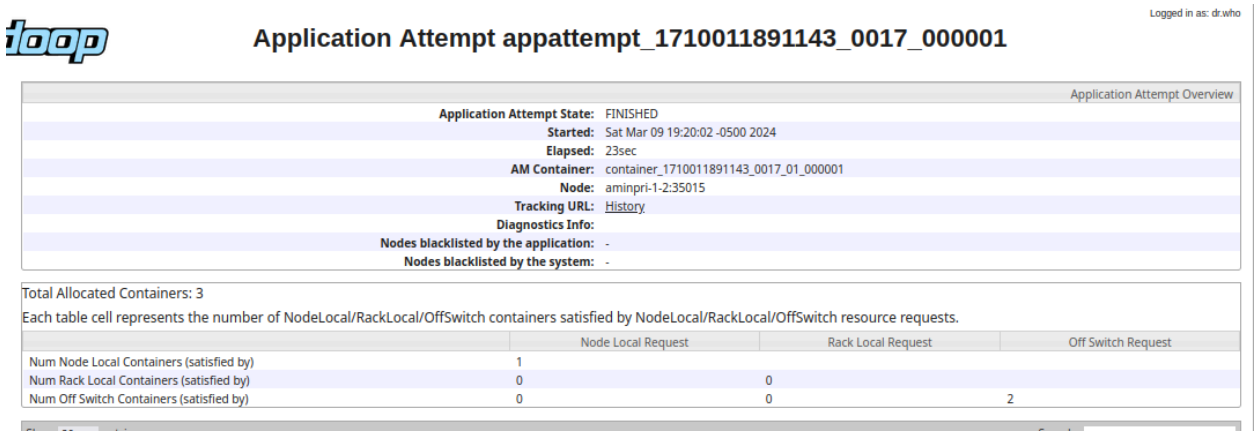
User:	hadoop
Name:	grep-sort
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sat Mar 09 19:20:02 -0500 2024
Launched:	Sat Mar 09 19:20:06 -0500 2024
Finished:	Sat Mar 09 19:20:26 -0500 2024
Elapsed:	24sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

Application Metrics

Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt:	0
Aggregate Resource Allocation:	61155 MB-seconds, 32 vcore-seconds
Aggregate Preempted Resource Allocation:	0 MB-seconds, 0 vcore-seconds

Show 20 entries Search:

Attempt ID	Started	Node	Logs	Nodes blacklisted by the app	Nodes blacklisted by the system
appattempt_1710011891143_0017_000001	Sat Mar 9 19:20:02 -0500 2024	http://aminpri-1-2:8042	Logs	0	0



Logged in as: dr.who

Application Attempt appattempt_1710011891143_0017_000001

Application Attempt Overview

Application Attempt State:	FINISHED
Started:	Sat Mar 09 19:20:02 -0500 2024
Elapsed:	23sec
AM Container:	container_1710011891143_0017_01_000001
Node:	aminpri-1-2:35015
Tracking URL:	History
Diagnostics Info:	
Nodes blacklisted by the application:	-
Nodes blacklisted by the system:	-

Total Allocated Containers: 3

Each table cell represents the number of NodeLocal/RackLocal/OffSwitch containers satisfied by NodeLocal/RackLocal/OffSwitch resource requests.

	Node Local Request	Rack Local Request	Off Switch Request
Num Node Local Containers (satisfied by)	1		
Num Rack Local Containers (satisfied by)	0	0	
Num Off Switch Containers (satisfied by)	0	0	2

Fair Scheduler Metrics for grep-search:
Application ID: application_1710011891143_0016
Elapsed Time: 3 minutes, 3 seconds
Total Allocated Containers: 29
Aggregate Resource Allocation: 997,877 MB-seconds, 772 vcore-seconds

Fair Scheduler Metrics for grep-sort:
Application ID: application_1710011891143_0017
Elapsed Time: 23 seconds
Total Allocated Containers: 3
Aggregate Resource Allocation: 61,155 MB-seconds, 32 vcore-seconds

Sort using fair scheduler:

hadoop@aminpri-1-2:~$ hadoop jar
$HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar \
-files /home/hadoop/assignment1/sort/mapper.py \
-input /assignment1/Input/access.log \
-output /assignment1/sort/sorted_output_fair \
-mapper "python3 /home/hadoop/assignment1/sort/mapper.py" \
-reducer "cat"

```
hadoop@aminpri-1-2:~$ hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar \
    -files /home/hadoop/assignment1/sort/mapper.py \
    -input /assignment1/Input/access.log \
    -output /assignment1/sort/sorted_output_fair \
    -mapper "python3 /home/hadoop/assignment1/sort/mapper.py" \
    -reducer "cat"
packageJobJar: [/tmp/hadoop-unjar4301506049257156198/] [] /tmp/streamjob1940140489495896597.jar tmpDir=null
2024-03-09 21:41:23,457 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-03-09 21:41:23,854 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-03-09 21:41:24,619 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging
/hadoop/.staging/job_1710011891143_0019
2024-03-09 21:41:25,144 INFO mapred.FileInputFormat: Total input files to process : 1
2024-03-09 21:41:25,175 INFO net.NetworkTopology: Adding a new node: /default-rack/127.0.0.1:9866
2024-03-09 21:41:25,294 INFO mapreduce.JobSubmitter: number of splits:26
2024-03-09 21:41:25,584 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1710011891143_0019
2024-03-09 21:41:25,585 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-03-09 21:41:26,155 INFO conf.Configuration: resource-types.xml not found
2024-03-09 21:41:26,156 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-03-09 21:41:26,269 INFO impl.YarnClientImpl: Submitted application application_1710011891143_0019
2024-03-09 21:41:26,345 INFO mapreduce.Job: The url to track the job: http://aminpri-1-2:8088/proxy/application_1710011
891143_0019/
2024-03-09 21:41:26,346 INFO mapreduce.Job: Running job: job_1710011891143_0019
2024-03-09 21:41:37,104 INFO mapreduce.Job: Job job_1710011891143_0019 running in uber mode : false
2024-03-09 21:41:37,111 INFO mapreduce.Job:  map 0% reduce 0%
2024-03-09 21:42:00,693 INFO mapreduce.Job:  map 1% reduce 0%
```

Metrics:

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers
application_1710011891143_0019	hadoop	streamjob1940140489495896597.jar	MAPREDUCE		default	0	Sat Mar 9 21:41:26 -0500 2024	Sat Mar 9 21:41:26 -0500 2024	Sat Mar 9 21:48:14 -0500 2024	FINISHED	SUCCEEDED	N/A

hadoop

Application application_1710011891143_0019

Application Overview

User:	hadoop
Name:	streamjob1940140489495896597.jar
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sat Mar 09 21:41:26 -0500 2024
Launched:	Sat Mar 09 21:41:26 -0500 2024
Finished:	Sat Mar 09 21:48:14 -0500 2024
Elapsed:	6mins, 48sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

Application Metrics

Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt:	0
Aggregate Resource Allocation:	2852619 MB-seconds, 2354 vcore-seconds
Aggregate Preempted Resource Allocation:	0 MB-seconds, 0 vcore-seconds

Show 20 entries Search:

Attempt ID	Started	Node	Logs	Nodes blacklisted by the app	Nodes blacklisted by the system
appattempt_1710011891143_0019_000001	Sat Mar 9 21:41:26 -0500 2024	http://aminpri-1-2:8042	Logs	0	0

Application Attempt appattempt_1710011891143_0019_000001

Application Attempt Overview

Application Attempt State:	FINISHED
Started:	Sat Mar 09 21:41:26 -0500 2024
Elapsed:	6mins, 48sec
AM Container:	container_1710011891143_0019_01_000001
Node:	aminpri-1-2:40083
Tracking URL:	History
Diagnostics Info:	
Nodes blacklisted by the application:	-
Nodes blacklisted by the system:	-

Total Allocated Containers: 29

Each table cell represents the number of NodeLocal/RackLocal/OffSwitch containers satisfied by NodeLocal/RackLocal/OffSwitch resource requests.

	Node Local Request	Rack Local Request	Off Switch Request
Num Node Local Containers (satisfied by)	26		
Num Rack Local Containers (satisfied by)	0	1	
Num Off Switch Containers (satisfied by)	0	0	2

Fair Scheduler Metrics:

Application ID: application_1710011891143_0019

Elapsed Time: 6 minutes, 48 seconds

Total Allocated Containers: 29

Aggregate Resource Allocation: 2,852,619 MB-seconds, 2,354 vcore-seconds

Analysis for wordcount:

Comparative Observations:

- **Execution Time:** The WordCount operation under the Capacity Scheduler was faster with an execution time of 4 minutes, 35 seconds compared to 7 minutes, 14 seconds under the Fair Scheduler.
- **Resource Allocation:** The Capacity Scheduler allocated resources more efficiently, with a lower aggregate resource allocation of 2,119,164 MB-seconds compared to the Fair Scheduler's 3,374,976 MB-seconds. The Capacity Scheduler used fewer vcore-seconds (1,772) compared to the Fair Scheduler (2,840).
- **Containers:** Both schedulers had a similar allocation with respect to node locality, with 28 node local containers for Capacity and 27 for Fair. Both schedulers had no rack local containers and had 2 off-switch containers.
- **Efficiency:** Based on the metrics, the Capacity Scheduler appears to be more efficient for the WordCount operation in terms of both execution time and resource utilization.
- **Scalability:** While both schedulers are handling the operation well without rack local containers, which suggests data locality is being effectively utilized, the Capacity Scheduler has shown better performance in this particular operation. The scalability in larger datasets or with different operations needs further testing.

This analysis suggests that for this particular WordCount operation, the Capacity Scheduler outperformed the Fair Scheduler in execution time and resource utilization.

Analysis for grep:

Comparative Observations:

- **Execution Time**: For grep-sort jobs, the Fair Scheduler (23 sec) was slightly slower than the Capacity Scheduler (19 sec). For grep-search jobs, the Fair Scheduler (3 mins, 3 sec) had a longer execution time compared to the Capacity Scheduler (2 mins, 13 sec).
- **Resource Allocation:** In terms of the grep-sort jobs, the Capacity Scheduler allocated a substantially higher amount of resources (406,676 MB-seconds) than the Fair Scheduler (61,155 MB-seconds). When looking at the grep-search jobs, the Fair Scheduler (997,877 MB-seconds) was more efficient than the Capacity Scheduler (1,026,215 MB-seconds), but not by a significant margin.
- **Container Utilization:** The grep-search jobs used a high number of containers for both schedulers, with the Capacity Scheduler using just one more container than the Fair Scheduler. The grep-sort jobs used the same number of containers (3) for both schedulers.
- **Efficiency**: The Fair Scheduler's grep-sort job was far more efficient in terms of resource allocation per second of execution time compared to the Capacity Scheduler. However, for the grep-search job, the Fair Scheduler and Capacity Scheduler showed similar resource utilization efficiency, with the Capacity Scheduler being slightly less efficient.

The Fair Scheduler is more efficient in resource allocation for grep-sort jobs, while the Capacity Scheduler leads in terms of execution time for both grep-sort and grep-search jobs.

Prithvi Amin (aminpri)
SP24-BL-ENGR-E516-14573

Analysis for sort:

Comparative Observations:

- **Execution Time:** The Sort operation under the Capacity Scheduler completed quicker with an execution time of 4 minutes, 54 seconds compared to 6 minutes, 48 seconds under the Fair Scheduler.
- **Resource Allocation:** The Capacity Scheduler utilized resources more efficiently, with a lower aggregate resource allocation of 1,944,639 MB-seconds compared to the Fair Scheduler's 2,852,619 MB-seconds. The Capacity Scheduler also used fewer vcore-seconds (1,584) compared to the Fair Scheduler (2,354).
- **Containers:** Both schedulers allocated the same total number of containers (29). However, the execution time and resource allocation efficiency were better with the Capacity Scheduler.
- **Efficiency:** Based on the available metrics, the Capacity Scheduler appears to be more efficient for the Sort job in terms of execution time and resource utilization.
- **Scalability:** Although data locality is equally good with both schedulers (26 node local containers for both), the Capacity Scheduler has demonstrated superior performance for this particular job. The scalability and performance with varying workloads or in a different cluster configuration would require further analysis.

This analysis suggests that for the Sort job provided, the Capacity Scheduler outperformed the Fair Scheduler in both execution time and resource allocation efficiency.