

# Digesh Patel , 202318038 , Hadoop 4 node cluster

Activities Brave Web Browser Feb 26 7:20 PM 39.8% 3.7% 0.0 B/s 42.0 B/s ~ (96%)

Namenode information All Applications Anime Dataset 2023 Instances | EC2 | ap-south-1 x Untitled document - Google +

WhatsApp

aws Services Search [Alt+S] Mumbai Digesh Patel

### Instances (1/4) Info

Find instance by attribute or tag (case-sensitive) Any state

Instance state = running X Clear filters

	Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone	Public IPv4 DNS
<input checked="" type="checkbox"/>	main	i-0b67c79280a97625b	Running	t2.medium	2/2 checks passed	View alarms +	ap-south-1b	ec2-43-204-217-33.ap-...
<input type="checkbox"/>	data1	i-0dc0ef51224a62c15	Running	t2.medium	2/2 checks passed	View alarms +	ap-south-1b	ec2-3-7-253-144.ap-so...
<input type="checkbox"/>	data2	i-04c02b6bc1da6394b	Running	t2.medium	2/2 checks passed	View alarms +	ap-south-1b	ec2-65-2-5-182.ap-sout...
<input type="checkbox"/>	data3	i-009350141e1f14785	Running	t2.medium	2/2 checks passed	View alarms +	ap-south-1b	ec2-3-108-238-145.ap-...

#### Instance: i-0b67c79280a97625b (main)

Details Status and alarms New Monitoring Security Networking Storage Tags

Instance summary Info

Instance ID i-0b67c79280a97625b (main)	Public IPv4 address 43.204.217.33 <a href="#">open address</a>	Private IPv4 addresses 172.31.6.131
---	---	--

CloudShell Feedback © 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Activities Terminal Feb 26 7:22 PM 39.5% 3.5% 90.1 B/s 140.1 B/s ~ (96%)

ubuntu@ip-172-31-6-131: ~/hadoop

```
ubuntu@ip-172-31-6-131:~/hadoop$ sbin/start-all.sh
This script is deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [ec2-43-204-217-33.ap-south-1.compute.amazonaws.com]
ec2-43-204-217-33.ap-south-1.compute.amazonaws.com: starting namenode, logging to /home/ubuntu/hadoop/logs/hadoop-ubuntu-namenode-ip-172-31-6-131.out
ec2-3-7-253-144.ap-south-1.compute.amazonaws.com: starting datanode, logging to /home/ubuntu/hadoop/logs/hadoop-ubuntu-datanode-ip-172-31-6-172.out
ec2-3-108-238-145.ap-south-1.compute.amazonaws.com: starting datanode, logging to /home/ubuntu/hadoop/logs/hadoop-ubuntu-datanode-ip-172-31-2-159.out
ec2-65-2-5-182.ap-south-1.compute.amazonaws.com: starting datanode, logging to /home/ubuntu/hadoop/logs/hadoop-ubuntu-datanode-ip-172-31-8-139.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/ubuntu/hadoop/logs/hadoop-ubuntu-secondarynamenode-ip-172-31-6-131.out
starting yarn daemons
starting resourcemanager, logging to /home/ubuntu/hadoop/logs/yarn-ubuntu-resourcemanager-ip-172-31-6-131.out
ec2-3-7-253-144.ap-south-1.compute.amazonaws.com: starting nodemanager, logging to /home/ubuntu/hadoop/logs/yarn-ubuntu-nodemanager-ip-172-31-6-172.out
ec2-3-108-238-145.ap-south-1.compute.amazonaws.com: starting nodemanager, logging to /home/ubuntu/hadoop/logs/yarn-ubuntu-nodemanager-ip-172-31-2-159.out
ec2-65-2-5-182.ap-south-1.compute.amazonaws.com: starting nodemanager, logging to /home/ubuntu/hadoop/logs/yarn-ubuntu-nodemanager-ip-172-31-8-139.out
ubuntu@ip-172-31-6-131:~/hadoop$ jps
15428 NameNode
15796 ResourceManager
16044 Jps
15663 SecondaryNameNode
ubuntu@ip-172-31-6-131:~/hadoop$
```

Activities Terminal Feb 26 7:22 PM 39.6% 5.3% 0.0 B/s 0.0 B/s ~ (96%)

ubuntu@ip-172-31-6-172: ~

```
ubuntu@ip-172-31-6-172:~$ jps
13346 Jps
13043 DataNode
13207 NodeManager
ubuntu@ip-172-31-6-172:~$
```

## About ratings csv :

columns : user\_id ,anime\_id , anime\_name , rating\_by\_user ,etc

35 million rows

Size : 4.2 GB

```
ubuntu@ip-172-31-6-131: ~  
ubuntu@ip-172-31-6-131:~$ ls  
full_ratings.csv  hadoop  mapper.py  reducer.py  sam.csv  
ubuntu@ip-172-31-6-131:~$ hdfs dfs -mkdir /input  
ubuntu@ip-172-31-6-131:~$ hdfs dfs -mkdir /output  
ubuntu@ip-172-31-6-131:~$ hdfs dfs -put ~/full_ratings.csv /input  
ubuntu@ip-172-31-6-131:~$ hdfs dfs -ls /input  
Found 1 items  
-rw-r--r--  3 ubuntu supergroup 4549801910 2024-02-26 13:57 /input/full_ratings.csv  
ubuntu@ip-172-31-6-131:~$
```

Activities Brave Web Browser Feb 26 7:29 PM 40.1% 2.3% 4.6 KB/s 432.4 B/s (96%)

Namenode information x All Applications Anime Dataset 2023 Instances | EC2 | ap-south-1 Untitled document - Google +

Not secure | http://ec2-43-204-217-33.ap-south-1.compute.amazonaws.com:50070/dfshealth.html#tab-overview

WhatsApp

Security is off.  
Safemode is off.  
4 files and directories, 34 blocks = 38 total filesystem object(s).  
Heap Memory used 49.68 MB of 294 MB Heap Memory. Max Heap Memory is 889 MB.  
Non Heap Memory used 47.69 MB of 48.5 MB Committed Non Heap Memory. Max Non Heap Memory is -1 B.

Configured Capacity:	43.07 GB
DFS Used:	12.81 GB (29.75%)
Non DFS Used:	8.7 GB
DFS Remaining:	21.56 GB (50.05%)
Block Pool Used:	12.81 GB (29.75%)
DataNodes usages% (Min/Median/Max/stdDev):	29.75% / 29.75% / 29.75% / 0.00%
Live Nodes	3 (Decommissioned: 0)
Dead Nodes	0 (Decommissioned: 0)
Decommissioning Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion	0

Activities Brave Web Browser Feb 26 7:29 PM 40.3% 5.4% 4.1 KB/s 5.1 KB/s ~ (96%)

Namenode Information x All Applications k Anime Dataset 2023 Instances | EC2 | ap-south-1 Untitled document - Google +

Not secure | http://ec2-43-204-217-33.ap-south-1.compute.amazonaws.com:50070/dfshealth.html#tab-datanode

WhatsApp

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

## Datanode Information

In operation

Node	Last contact	Admin State	Capacity	Used	Non DFS Used	Remaining	Blocks	Block pool used	Failed Volumes	Version
ip-172-31-8-139.ap-south-1.compute.internal:50010 (172.31.8.139:50010)	2	In Service	14.36 GB	4.27 GB	2.9 GB	7.19 GB	34	4.27 GB (29.75%)	0	2.7.3
ip-172-31-6-172.ap-south-1.compute.internal:50010 (172.31.6.172:50010)	1	In Service	14.36 GB	4.27 GB	2.9 GB	7.19 GB	34	4.27 GB (29.75%)	0	2.7.3
ip-172-31-2-159.ap-south-1.compute.internal:50010 (172.31.2.159:50010)	2	In Service	14.36 GB	4.27 GB	2.9 GB	7.19 GB	34	4.27 GB (29.75%)	0	2.7.3

Decommissioning

## mapper.py

```

ubuntu@ip-172-31-6-131: ~ x ubuntu@ip-172-31-6-172: ~
#!/usr/bin/python3

import sys

# Mapper function
for line in sys.stdin:

    # Skip first row as it contains column names
    if line.startswith("username"):
        continue

    # Split the line into fields
    fields = line.strip().split(',')

    # getting variables
    try:
        anime_id = fields[1] # index for anime ID
        rating = fields[2] # index for rating
        anime_name = fields[5] # index for anime name
    except:
        continue

    # Emit key-value pair (anime id, rating, anime name)
    print('%s\t%s\t%s' % (anime_id, rating, anime_name))

```

## reducer.py

```
ubuntu@ip-172-31-6-131: ~ × ubuntu@ip-172-31-6-172: ~ × ubuntu@ip-172-31-8-139: ~/hadoop/et... ×
#!/usr/bin/python3
import sys

# Initialize variables to store intermediate values
current_anime_id = None
current_anime_name = None
current_total_score = 0
current_rating_count = 0

# Reducer function
for line in sys.stdin:
    # Strip and split the input
    anime_id, rating, anime_name = line.strip().split('\t')

    # Convert rating to integer
    rating = int(rating)

    # Check if the anime ID has changed
    if current_anime_id != anime_id:
        # If this is not the first anime ID, emit the result for the previous one
        if current_anime_id:
            # Calculate average score
            avg_score = current_total_score / current_rating_count
            # Emit the result
            print('ID: %s\tName: %s\tAvgScore: %.2f\tCounts: %s' % (current_anime_id, current_anime_name, avg_score, current_rating_count))

        # Reset variables for the new anime ID
        current_anime_id = anime_id
        current_anime_name = anime_name
        current_total_score = 0
        current_rating_count = 0

    # Accumulate total score and count for the current anime ID
    current_total_score += rating
    current_rating_count += 1

# Emit the result for the last anime ID
if current_anime_id:
    # Calculate average score
    avg_score = current_total_score / current_rating_count
    # Emit the result
    print('ID: %s\tName: %s\tAvgScore: %.2f\tCounts: %s' % (current_anime_id, current_anime_name, avg_score, current_rating_count))
```

# Running Hadoop Job and show output

```
Activities Terminal Feb 26 7:47 PM 41.0% 3.1% 94.0 B/s 166.0 B/s ~ (96%)

ubuntu@ip-172-31-6-131: ~
ubuntu@ip-172-31-6-131: ~ x ubuntu@ip-172-31-6-172: ~ x ubuntu@ip-172-31-8-139: ~/hadoop/et... x ubuntu@ip-172-31-2-159: ~/hadoop/et... x
ubuntu@ip-172-31-6-131:~$ hadoop jar /home/ubuntu/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.3.jar -file /home/ubuntu/mapper.py -mapper mappe
r.py -file /home/ubuntu/reducer.py -reducer reducer.py -input /input/full_ratings.csv -output /output/output-10
24/02/26 14:16:04 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/ubuntu/mapper.py, /home/ubuntu/reducer.py, /tmp/hadoop-unjar4944542782034415697/] [] /tmp/streamjob109097731097890992.jar tmpD
ir=null
24/02/26 14:16:05 INFO client.RMProxy: Connecting to ResourceManager at ec2-43-204-217-33.ap-south-1.compute.amazonaws.com/172.31.6.131:8032
24/02/26 14:16:05 INFO client.RMProxy: Connecting to ResourceManager at ec2-43-204-217-33.ap-south-1.compute.amazonaws.com/172.31.6.131:8032
24/02/26 14:16:06 INFO mapred.FileInputFormat: Total input paths to process : 1
24/02/26 14:16:06 INFO mapreduce.JobSubmitter: number of splits:34
24/02/26 14:16:06 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1708955458547_0003
24/02/26 14:16:06 INFO impl.YarnClientImpl: Submitted application application_1708955458547_0003
24/02/26 14:16:06 INFO mapreduce.Job: The url to track the job: http://ec2-43-204-217-33.ap-south-1.compute.amazonaws.com:8088/proxy/application_1708
955458547_0003/
24/02/26 14:16:06 INFO mapreduce.Job: Running job: job_1708955458547_0003
24/02/26 14:16:12 INFO mapreduce.Job: Job job_1708955458547_0003 running in uber mode : false
24/02/26 14:16:12 INFO mapreduce.Job: map 0% reduce 0%
24/02/26 14:16:32 INFO mapreduce.Job: map 3% reduce 0%
24/02/26 14:16:35 INFO mapreduce.Job: map 6% reduce 0%
24/02/26 14:16:36 INFO mapreduce.Job: map 7% reduce 0%
24/02/26 14:16:37 INFO mapreduce.Job: map 9% reduce 0%
24/02/26 14:16:38 INFO mapreduce.Job: map 12% reduce 0%
24/02/26 14:16:39 INFO mapreduce.Job: map 13% reduce 0%
24/02/26 14:16:40 INFO mapreduce.Job: map 15% reduce 0%
24/02/26 14:16:41 INFO mapreduce.Job: map 19% reduce 0%
24/02/26 14:16:42 INFO mapreduce.Job: map 20% reduce 0%
24/02/26 14:16:43 INFO mapreduce.Job: map 22% reduce 0%
24/02/26 14:16:44 INFO mapreduce.Job: map 25% reduce 0%
24/02/26 14:16:45 INFO mapreduce.Job: map 26% reduce 0%
24/02/26 14:16:46 INFO mapreduce.Job: map 28% reduce 0%
24/02/26 14:16:47 INFO mapreduce.Job: map 29% reduce 0%
24/02/26 14:16:48 INFO mapreduce.Job: map 31% reduce 0%
24/02/26 14:16:49 INFO mapreduce.Job: map 35% reduce 0%
24/02/26 14:16:50 INFO mapreduce.Job: map 40% reduce 0%
24/02/26 14:16:51 INFO mapreduce.Job: map 42% reduce 0%
24/02/26 14:16:52 INFO mapreduce.Job: map 44% reduce 0%
24/02/26 14:16:54 INFO mapreduce.Job: map 46% reduce 0%
24/02/26 14:16:55 INFO mapreduce.Job: map 48% reduce 0%
```

```
Activities Terminal Feb 26 7:51 PM 40.6% 2.3% 188.0 B/s 132.0 B/s ~ (96%)

ubuntu@ip-172-31-6-131: ~
ubuntu@ip-172-31-6-131: ~ x ubuntu@ip-172-31-6-172: ~ x ubuntu@ip-172-31-8-139: ~/hadoop/et... x ubuntu@ip-172-31-2-159: ~/hadoop/et... x
ubuntu@ip-172-31-6-131:~$ hdfs dfs -ls /output/10
ls: '/output/10': No such file or directory
ubuntu@ip-172-31-6-131:~$ hdfs dfs -ls /output/output-10
Found 2 items
-rw-r--r-- 3 ubuntu supergroup 0 2024-02-26 14:18 /output/output-10/_SUCCESS
-rw-r--r-- 3 ubuntu supergroup 609389 2024-02-26 14:18 /output/output-10/part-00000
ubuntu@ip-172-31-6-131:~$ hdfs dfs -cat /output/output-10/part-00000
ID: 1 Name: Cowboy Bebop AvgScore: 5.79 Counts: 51707
ID: 100 Name: Shin Shirayuki-hime Densetsu Prêtear AvgScore: 4.17 Counts: 7569
ID: 1000 Name: Uchuu Kaizoku Captain Herlock AvgScore: 2.68 Counts: 3000
ID: 10003 Name: Kago Shintarou Anime Sakuhin Shuu AvgScore: 2.15 Counts: 457
ID: 1001 Name: Tide-Line Blue Special AvgScore: 3.87 Counts: 368
ID: 10012 Name: Carnival Phantasm AvgScore: 4.52 Counts: 12665
ID: 10013 Name: Shouwa Monogatari (Movie) AvgScore: 2.21 Counts: 240
ID: 10014 Name: Shouwa Monogatari AvgScore: 2.15 Counts: 1000
ID: 10015 Name: Yu-Gi-Oh! Zexal AvgScore: 3.65 Counts: 3083
ID: 10016 Name: Kizuna Ichigeki AvgScore: 4.71 Counts: 1152
ID: 10017 Name: Dragon Ball: Super Salya-jin Zetsumetsu Keikaku AvgScore: 4.76 Counts: 2537
ID: 1002 Name: Top wo Nerae 2! Diebuster AvgScore: 4.54 Counts: 6859
ID: 10020 Name: Ore no Imouto ga Konnani Kawaii Wake ga Nai Specials AvgScore: 6.03 Counts: 18009
ID: 10029 Name: Coquelicot-zaka kara AvgScore: 5.01 Counts: 8661
ID: 1003 Name: Aa! Megami-san! (TV) Specials AvgScore: 5.10 Counts: 4312
ID: 10030 Name: Bakuman. 2nd Season AvgScore: 6.00 Counts: 20381
ID: 10033 Name: Toriko AvgScore: 3.46 Counts: 7637
ID: 10036 Name: Boku no Chikyuu wo Mamotte: Kiniro no Toki Nasarete AvgScore: 1.78 Counts: 152
ID: 1004 Name: Kanojo to Kanojo no Neko AvgScore: 5.49 Counts: 12252
ID: 10043 Name: Sailor Fuku Shinryou Tsunaka AvgScore: 2.81 Counts: 278
```


# Execution time

Activities Brave Web Browser Feb 26 7:51 PM 41.5% 3.6% 0.0 B/s 0.0 B/s ~ (96%)

Browsing HDFS Application application\_ x Anime Dataset 2023 Instances | EC2 | ap-south-1 Untitled document - Google +

Not secure | http://ec2-43-204-217-33.ap-south-1.compute.amazonaws.com:8088/cluster/app/application\_1708955458547\_0003

WhatsApp



## Application application\_1708955458547\_0003

Logged in as: dr.who

Cluster

About

Nodes

Node Labels

Applications

NEW

NEW SAVING

SUBMITTED

ACCEPTED

RUNNING

FINISHED

FAILED

KILLED

Scheduler

Kill Application

Application Overview

User: ubuntu

Name: streamjob1090977331097890992.jar

Application Type: MAPREDUCE

Application Tags:

YarnApplicationState: FINISHED

Queue: default

FinalStatus Reported by AM: SUCCEEDED

Started: Mon Feb 26 14:16:06 +0000 2024

Elapsed: 2mins, 23sec

Tracking URL: History

Diagnostics:

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: 1919309 MB-seconds, 1704 vcore-seconds

Show 20 entries

Search:

Attempt ID	Started	Node	Logs	Blacklisted Nodes
appattempt_1708955458547_0003_000001	Mon Feb 26 19:46:06 +0550 2024	http://ig-172-31-2-159.ap-south-1.compute.internal:8042	Logs	N/A

Showing 1 to 1 of 1 entries

First Previous 1 Next Last