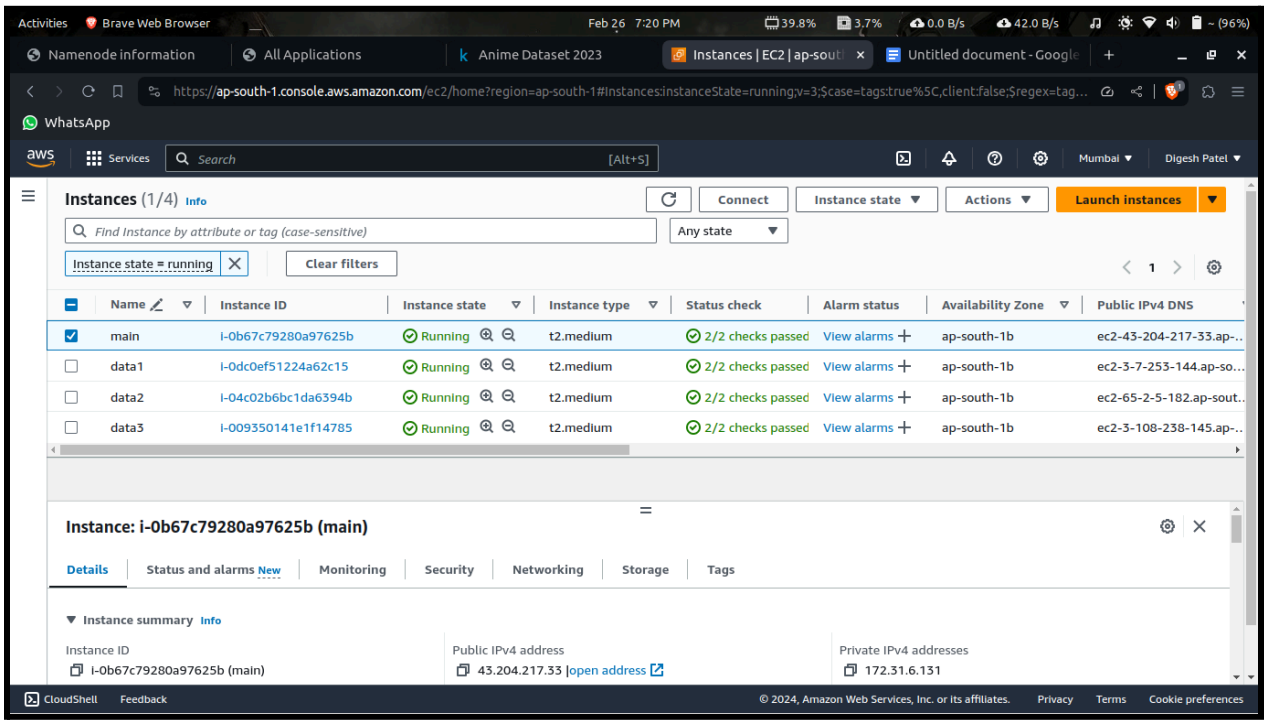
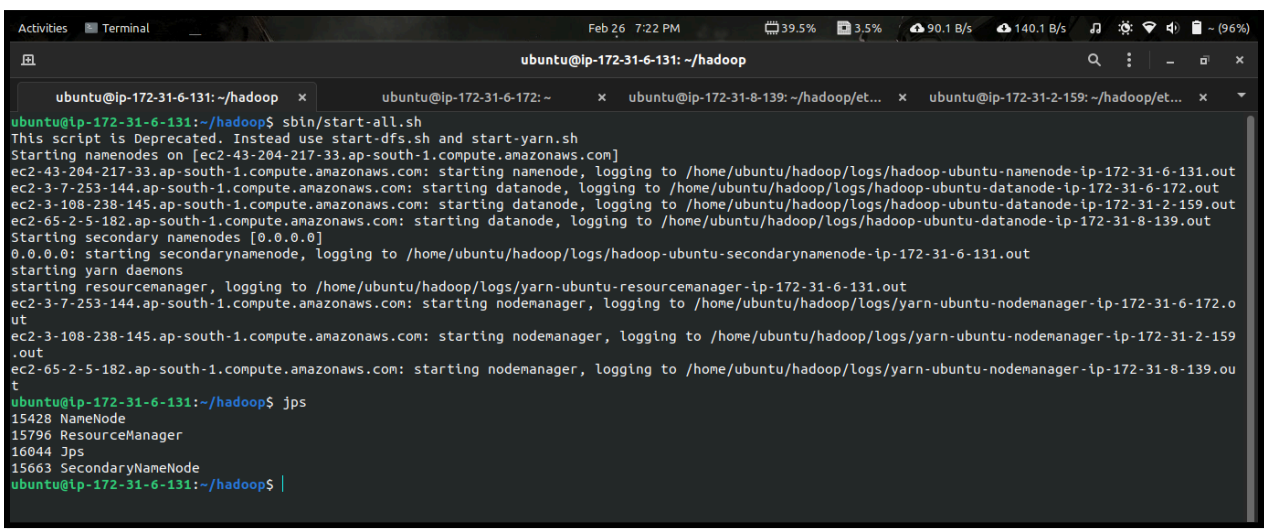
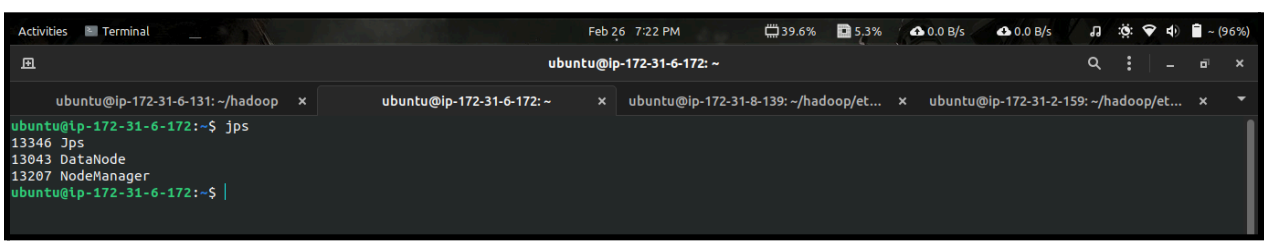
Digesh Patel , 202318038 , Hadoop 4 node cluster

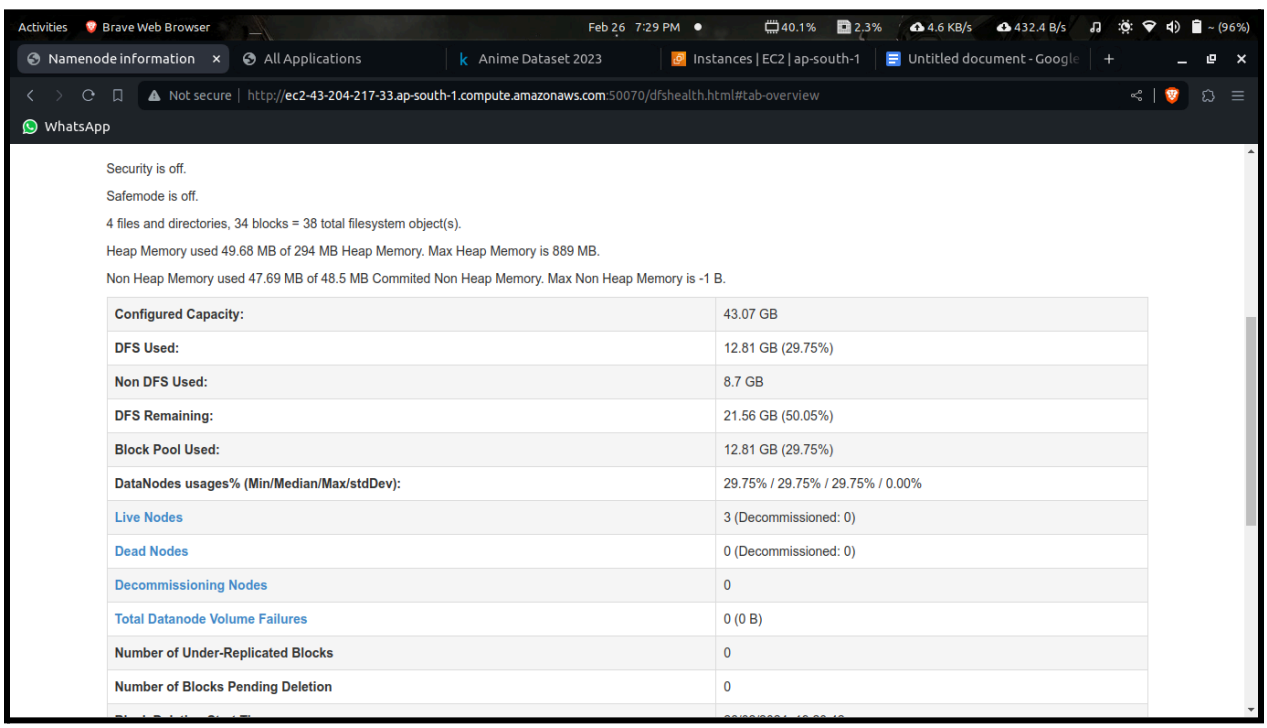


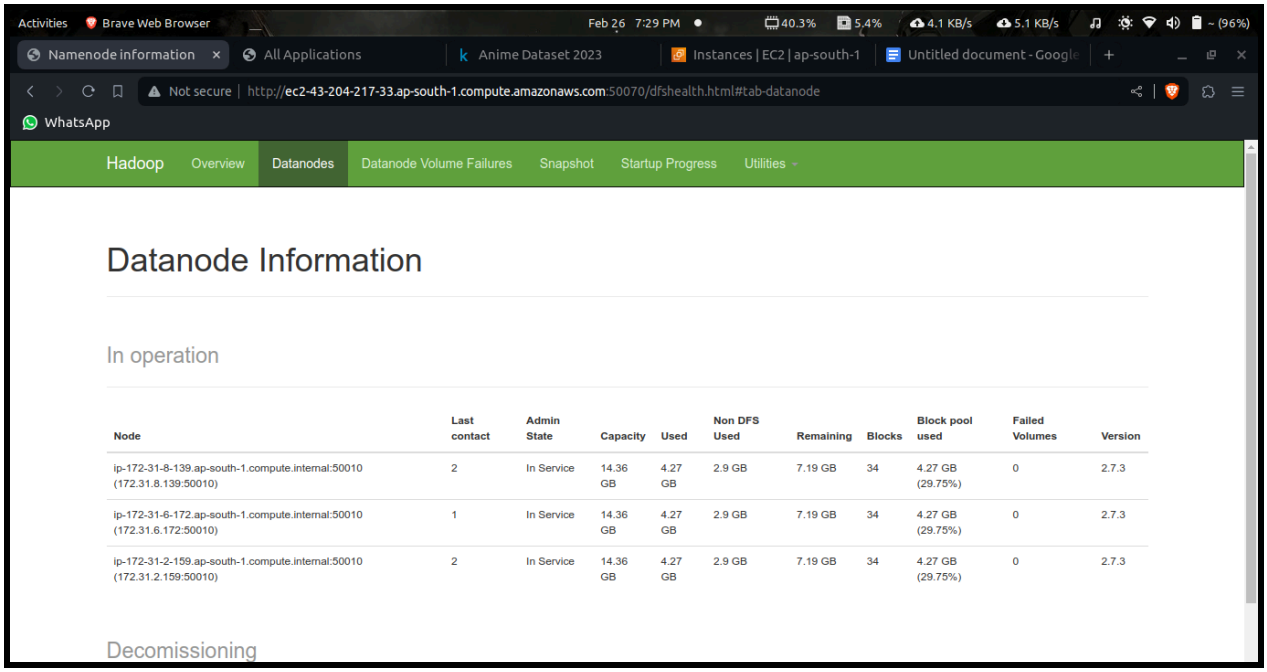




About ratings csv :

columns : user_id ,anime_id , anime_name , rating_by_user ,etc
35 million rows
Size : 4.2 GB





mapper.py

ubuntu@ip-172-31-6-131: ~ | ubuntu@ip-172-31-6-172: ~

```python
#!/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...

```python
#!/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

```
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/streamjob1090977331097890992.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%
```

```
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=4549937078
        File Output Format Counters
                Bytes Written=609389
24/02/26 14:18:31 INFO streaming.StreamJob: Output directory: /output/output-10
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 Prétear      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 Saiya-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-sama! (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 Tsumaka      AvgScore: 2.81  Counts: 278
```

Execution time