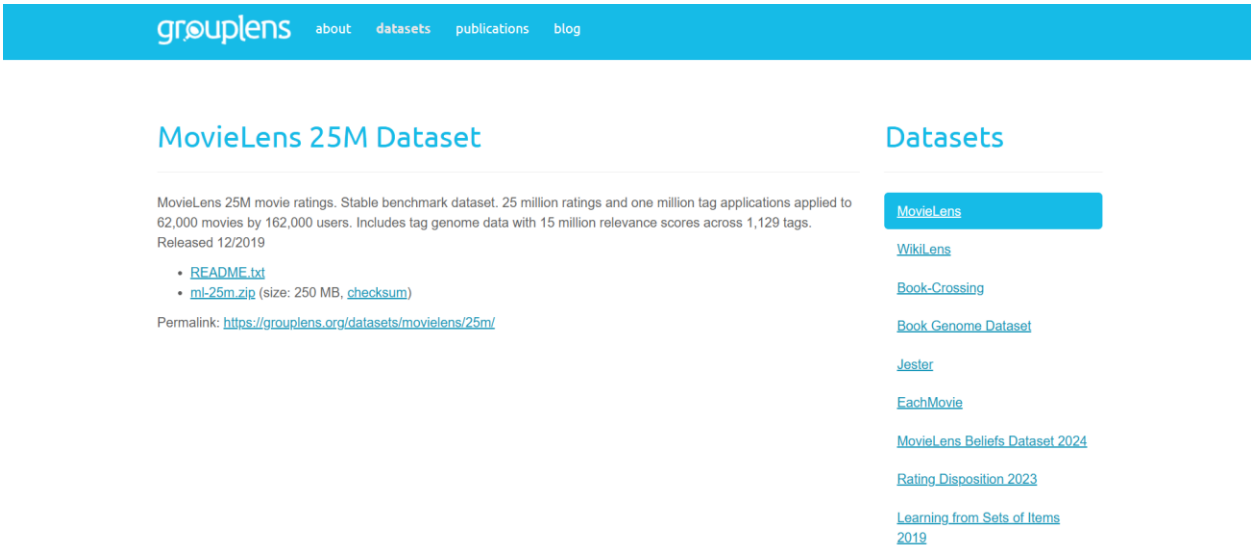
🎬 Hadoop-Powered Movies Analytics

Introduction

This project uses the MapReduce programming model to analyze and process the MovieLens dataset in a distributed environment using Hadoop. The main objective is to extract meaningful insights from user ratings data—specifically, to compute the average ratings of movies and retrieve the top 20 highest-rated movies along with their titles. The analysis is conducted in two phases (MapReduce 1 and MapReduce 2), processing two datasets: `ratings.csv` (user ratings) and `movies.csv` (movie metadata)



mapreduce1

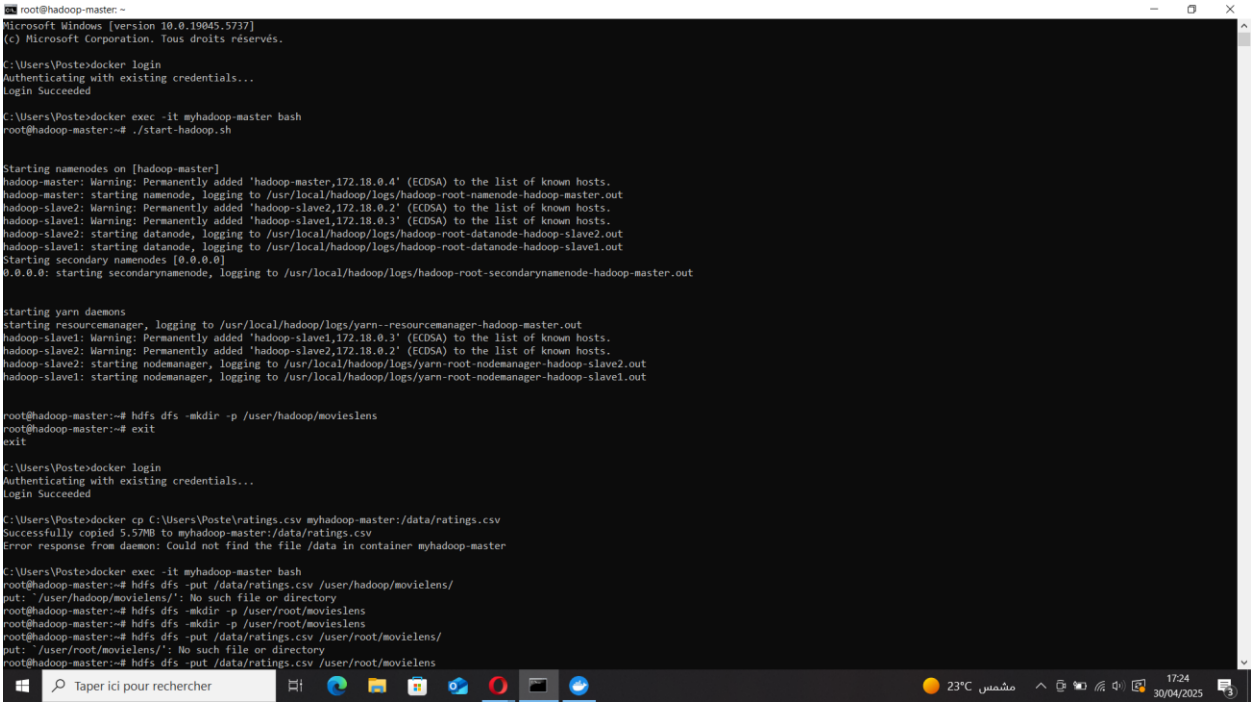
In the first phase, MapReduce 1, the input data consists of the `ratings.csv` file located at `/user/root/input/ratings.csv`. Each record in this file follows the format: `userId,movieId,rating,timestamp`. The Mapper reads each line, extracts the `movieId` and `rating`, and emits a key-value pair where the key is the movie ID and the value is the rating. The Reducer then receives all ratings for each movie ID, computes their average, and outputs a result in the format: `movieId average_rating`. This output is stored in the HDFS directory `/user/hadoop/output/part-00000`.

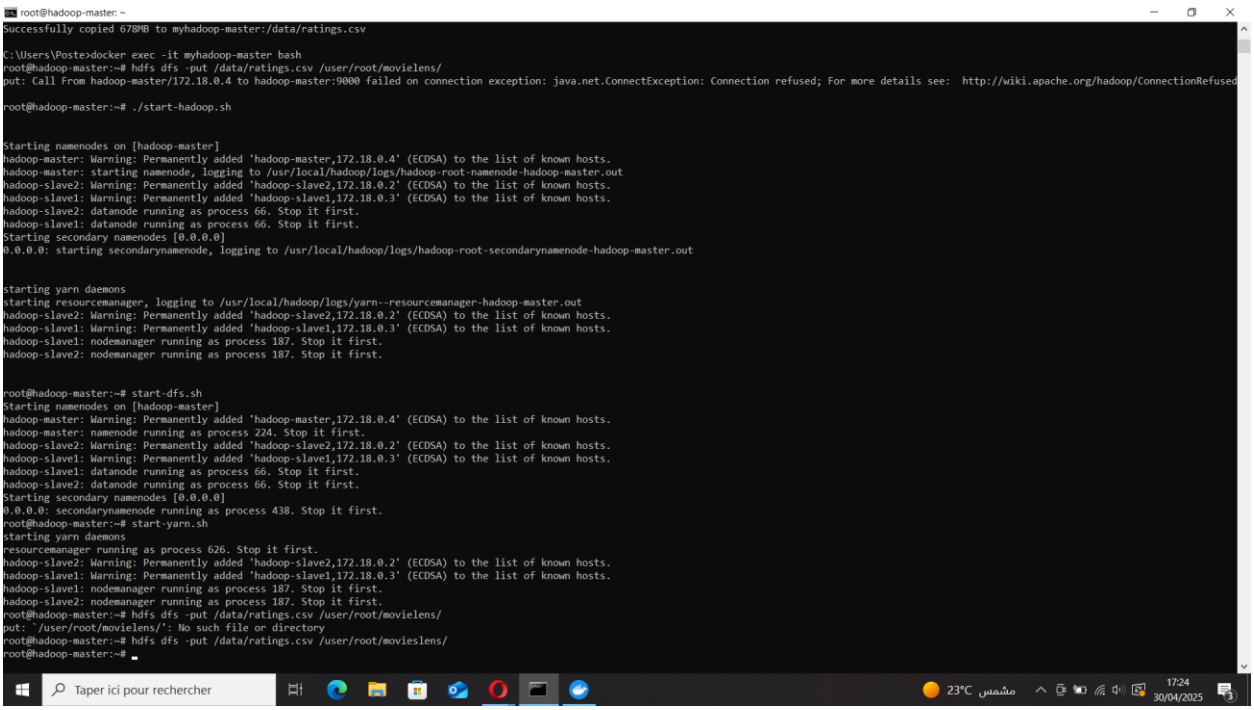
ratings.csv

```
userId,movieId,rating,timestamp
1,296,5.0,1147880044
1,306,3.5,1147868817
1,307,5.0,1147868828
1,665,5.0,1147878820
1,899,3.5,1147868510
1,1088,4.0,1147868495
1,1175,3.5,1147868826
1,1217,3.5,1147878326
1,1237,5.0,1147868839
1,1250,4.0,1147868414
1,1260,3.5,1147877857
1,1653,4.0,1147868097
1,2011,2.5,1147868079
1,2012,2.5,1147868068
1,2068,2.5,1147869044
1,2161,3.5,1147868609
```

movies.csv

```
movieId,title,genres
1,Toy Story (1995),Adventure|Animation|Children|Comedy|Fantasy
2,Jumanji (1995),Adventure|Children|Fantasy
3,Grumpier Old Men (1995),Comedy|Romance
4,Waiting to Exhale (1995),Comedy|Drama|Romance
5,Father of the Bride Part II (1995),Comedy
6,Heat (1995),Action|Crime|Thriller
7,Sabrina (1995),Comedy|Romance
8,Tom and Huck (1995),Adventure|Children
9,Sudden Death (1995),Action
10,GoldenEye (1995),Action|Adventure|Thriller
11,"American President, The (1995)",Comedy|Drama|Romance
12,Dracula: Dead and Loving It (1995),Comedy|Horror
13,Balto (1995),Adventure|Animation|Children
14,Nixon (1995),Drama
15,Cutthroat Island (1995),Action|Adventure|Romance
16,Casino (1995),Crime|Drama
17,Sense and Sensibility (1995),Drama|Romance
18,Four Rooms (1995),Comedy
19,Ace Ventura: When Nature Calls (1995),Comedy
```


root@hadoop-master: ~
Microsoft Windows [version 10.0.19045.5737]
(c) Microsoft Corporation. Tous droits réservés.
C:\Users\Poste>docker login
Authenticating with existing credentials...
Login Succeeded
C:\Users\Poste>docker exec -it myhadoop-master bash
root@hadoop-master:~# ./start-hadoop.sh
Starting namenodes on [hadoop-master]
hadoop-master: Warning: Permanently added 'hadoop-master,172.18.0.4' (ECDSA) to the list of known hosts.
hadoop-master: starting namenode, logging to /usr/local/hadoop/logs/hadoop-root-namenode-hadoop-master.out
hadoop-slave2: Warning: Permanently added 'hadoop-slave2,172.18.0.2' (ECDSA) to the list of known hosts.
hadoop-slave1: Warning: Permanently added 'hadoop-slave1,172.18.0.3' (ECDSA) to the list of known hosts.
hadoop-slave2: starting datanode, logging to /usr/local/hadoop/logs/hadoop-root-datanode-hadoop-slave2.out
hadoop-slave1: starting datanode, logging to /usr/local/hadoop/logs/hadoop-root-datanode-hadoop-slave1.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-root-secondarynamenode-hadoop-master.out
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn--resourcemanager-hadoop-master.out
hadoop-slave1: Warning: Permanently added 'hadoop-slave1,172.18.0.3' (ECDSA) to the list of known hosts.
hadoop-slave2: Warning: Permanently added 'hadoop-slave2,172.18.0.2' (ECDSA) to the list of known hosts.
hadoop-slave2: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-root-nodemanager-hadoop-slave2.out
hadoop-slave1: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-root-nodemanager-hadoop-slave1.out
root@hadoop-master:~# hdfs dfs -mkdir -p /user/hadoop/movieslens
root@hadoop-master:~# exit
exit
C:\Users\Poste>docker login
Authenticating with existing credentials...
Login Succeeded
C:\Users\Poste>docker cp C:\Users\Poste\ratings.csv myhadoop-master:/data/ratings.csv
Successfully copied 5.57MB to myhadoop-master:/data/ratings.csv
Error response from daemon: Could not find the file /data in container myhadoop-master
C:\Users\Poste>docker exec -it myhadoop-master bash
root@hadoop-master:~# hdfs dfs -put /data/ratings.csv /user/hadoop/movielens/
put: `/user/hadoop/movielens/': No such file or directory
root@hadoop-master:~# hdfs dfs -mkdir -p /user/root/movieslens
root@hadoop-master:~# hdfs dfs -mkdir -p /user/root/movieslens
root@hadoop-master:~# hdfs dfs -put /data/ratings.csv /user/root/movielens/
put: `/user/root/movielens/': No such file or directory
root@hadoop-master:~# hdfs dfs -put /data/ratings.csv /user/root/movielens
Taper ici pour rechercher
23°C مشمس
17:24
30/04/2025


root@hadoop-master: ~
Successfully copied 678MB to myhadoop-master:/data/ratings.csv
C:\Users\Poste>docker exec -it myhadoop-master bash
root@hadoop-master:~# hdfs dfs -put /data/ratings.csv /user/root/movielens/
put: Call From hadoop-master/172.18.0.4 to hadoop-master:9000 failed on connection exception: java.net.ConnectException: Connection refused; For more details see: http://wiki.apache.org/hadoop/ConnectionRefused
root@hadoop-master:~# ./start-hadoop.sh
Starting namenodes on [hadoop-master]
hadoop-master: Warning: Permanently added 'hadoop-master,172.18.0.4' (ECDSA) to the list of known hosts.
hadoop-master: starting namenode, logging to /usr/local/hadoop/logs/hadoop-root-namenode-hadoop-master.out
hadoop-slave2: Warning: Permanently added 'hadoop-slave2,172.18.0.2' (ECDSA) to the list of known hosts.
hadoop-slave1: Warning: Permanently added 'hadoop-slave1,172.18.0.3' (ECDSA) to the list of known hosts.
hadoop-slave2: datanode running as process 66. Stop it first.
hadoop-slave1: datanode running as process 66. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-root-secondarynamenode-hadoop-master.out
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn--resourcemanager-hadoop-master.out
hadoop-slave2: Warning: Permanently added 'hadoop-slave2,172.18.0.2' (ECDSA) to the list of known hosts.
hadoop-slave1: Warning: Permanently added 'hadoop-slave1,172.18.0.3' (ECDSA) to the list of known hosts.
hadoop-slave1: nodemanager running as process 187. Stop it first.
hadoop-slave2: nodemanager running as process 187. Stop it first.
root@hadoop-master:~# start-dfs.sh
Starting namenodes on [hadoop-master]
hadoop-master: Warning: Permanently added 'hadoop-master,172.18.0.4' (ECDSA) to the list of known hosts.
hadoop-master: namenode running as process 224. Stop it first.
hadoop-slave2: Warning: Permanently added 'hadoop-slave2,172.18.0.2' (ECDSA) to the list of known hosts.
hadoop-slave1: Warning: Permanently added 'hadoop-slave1,172.18.0.3' (ECDSA) to the list of known hosts.
hadoop-slave1: datanode running as process 66. Stop it first.
hadoop-slave2: datanode running as process 66. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 438. Stop it first.
root@hadoop-master:~# start-yarn.sh
starting yarn daemons
resourcemanager running as process 626. Stop it first.
hadoop-slave2: Warning: Permanently added 'hadoop-slave2,172.18.0.2' (ECDSA) to the list of known hosts.
hadoop-slave1: Warning: Permanently added 'hadoop-slave1,172.18.0.3' (ECDSA) to the list of known hosts.
hadoop-slave1: nodemanager running as process 187. Stop it first.
hadoop-slave2: nodemanager running as process 187. Stop it first.
root@hadoop-master:~# hdfs dfs -put /data/ratings.csv /user/root/movielens/
put: `/user/root/movielens/': No such file or directory
root@hadoop-master:~# hdfs dfs -put /data/ratings.csv /user/root/movieslens/
root@hadoop-master:~#
Taper ici pour rechercher
23°C مشمس
17:24
30/04/2025

Containers / myhadoop-master

myhadoop-master

9570146df163 soukiolfa92/hadoop:hv-2.7.2

16010:16010 50070:50070 7077:7077 8088:8088 Show less

Logs Inspect Bind mounts Exec **Files** Stats

Name ↑

- root
  - .bash_history
  - .bashrc

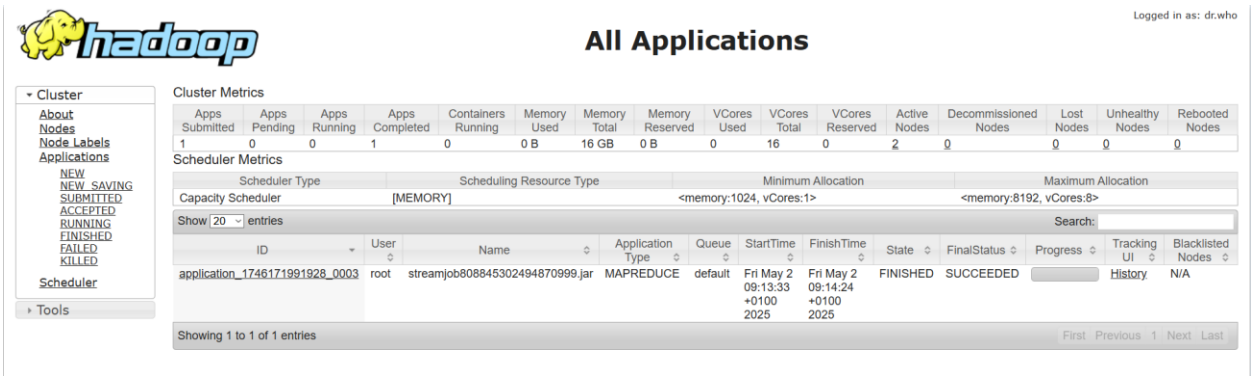
/data/mapper.py

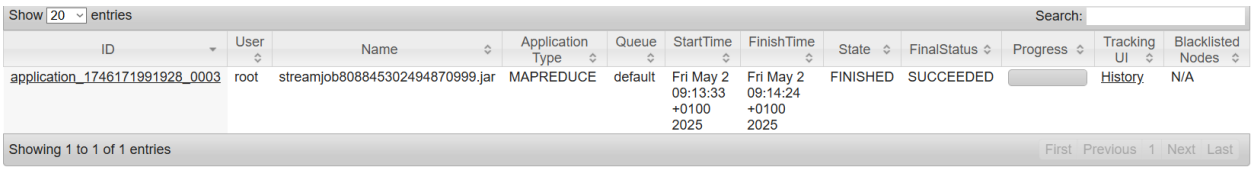
```
1  #!/usr/bin/env python3
2  import sys
3
4  for line in sys.stdin:
5      parts = line.strip().split(",")
6      if len(parts) != 4 or parts[0] == "userId":
7          continue
8      movie_id = parts[1]
9      rating = parts[2]
10     print(f"{movie_id}\t{rating}")
11
```

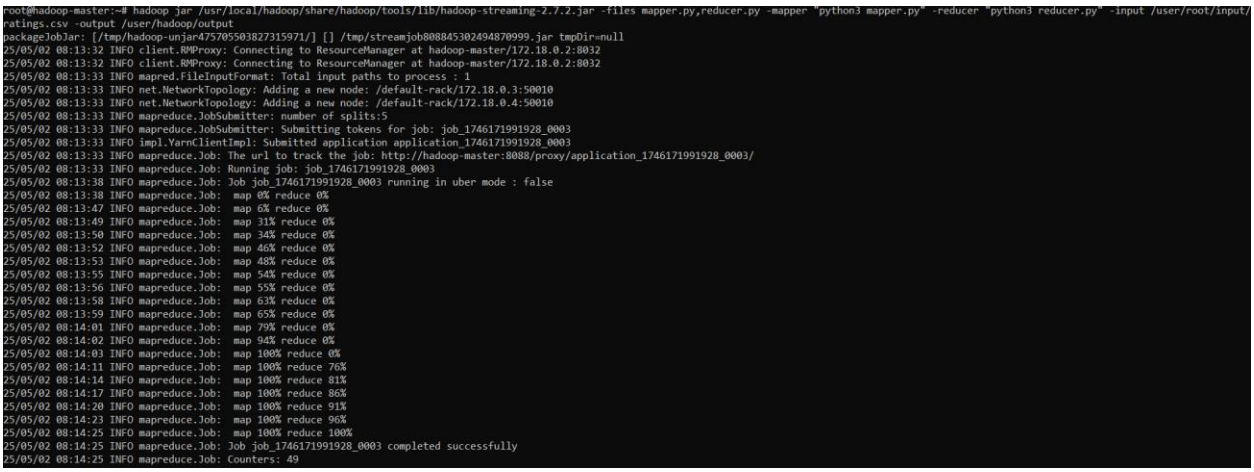
reducer.py > ...

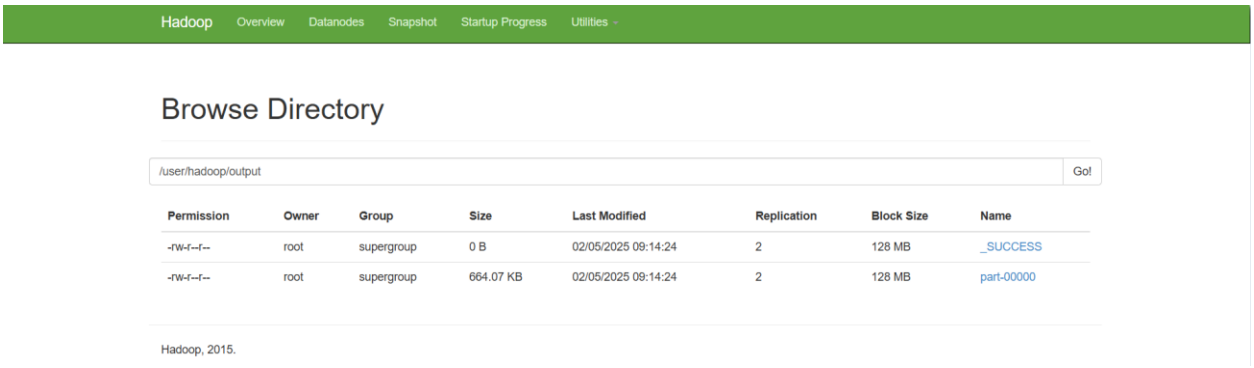
```python
#!/usr/bin/env python3
import sys

current_movie = None
total_rating = 0.0
count = 0

for line in sys.stdin:
    movie_id, rating = line.strip().split("\t")
    if current_movie == movie_id:
        total_rating += float(rating)
        count += 1
    else:
        if current_movie:
            average = total_rating / count
            print(f"{current_movie}\t{average:.2f}")
        current_movie = movie_id
        total_rating = float(rating)
        count = 1

if current_movie == movie_id:
    average = total_rating / count
    print(f"{current_movie}\t{average:.2f}")
```

```
root@hadoop-master:~# hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar -files mapper.py,reducer.py -mapper "python3 mapper.py" -reducer "python3 reducer.py" -input /user/root/input/
ratings.csv -output /user/hadoop/output
packageJobJar: [/tmp/hadoop-unjar4757055038273159 71/] [] /tmp/streamjob8088453024948709 99.jar tmpDir=null
25/05/02 08:13:32 INFO client.RMProxy: Connecting to ResourceManager at hadoop-master/172.18.0.2:8032
25/05/02 08:13:32 INFO client.RMProxy: Connecting to ResourceManager at hadoop-master/172.18.0.2:8032
25/05/02 08:13:33 INFO mapred.FileInputFormat: Total input paths to process : 1
25/05/02 08:13:33 INFO net.NetworkTopology: Adding a new node: /default-rack/172.18.0.3:50010
25/05/02 08:13:33 INFO net.NetworkTopology: Adding a new node: /default-rack/172.18.0.4:50010
25/05/02 08:13:33 INFO mapreduce.JobSubmitter: number of splits:5
25/05/02 08:13:33 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1746171991928_0003
25/05/02 08:13:33 INFO impl.YarnClientImpl: Submitted application application_1746171991928_0003
25/05/02 08:13:33 INFO mapreduce.Job: The url to track the job: http://hadoop-master:8088/proxy/application_1746171991928_0003/
25/05/02 08:13:33 INFO mapreduce.Job: Running job: job_1746171991928_0003
25/05/02 08:13:38 INFO mapreduce.Job: Job job_1746171991928_0003 running in uber mode : false
25/05/02 08:13:38 INFO mapreduce.Job:  map 0% reduce 0%
25/05/02 08:13:47 INFO mapreduce.Job:  map 6% reduce 0%
25/05/02 08:13:49 INFO mapreduce.Job:  map 31% reduce 0%
25/05/02 08:13:50 INFO mapreduce.Job:  map 34% reduce 0%
25/05/02 08:13:52 INFO mapreduce.Job:  map 46% reduce 0%
25/05/02 08:13:53 INFO mapreduce.Job:  map 48% reduce 0%
25/05/02 08:13:55 INFO mapreduce.Job:  map 54% reduce 0%
25/05/02 08:13:56 INFO mapreduce.Job:  map 55% reduce 0%
25/05/02 08:13:58 INFO mapreduce.Job:  map 63% reduce 0%
25/05/02 08:13:59 INFO mapreduce.Job:  map 65% reduce 0%
25/05/02 08:14:01 INFO mapreduce.Job:  map 79% reduce 0%
25/05/02 08:14:02 INFO mapreduce.Job:  map 94% reduce 0%
25/05/02 08:14:03 INFO mapreduce.Job:  map 100% reduce 0%
25/05/02 08:14:11 INFO mapreduce.Job:  map 100% reduce 76%
25/05/02 08:14:14 INFO mapreduce.Job:  map 100% reduce 81%
25/05/02 08:14:17 INFO mapreduce.Job:  map 100% reduce 86%
25/05/02 08:14:20 INFO mapreduce.Job:  map 100% reduce 91%
25/05/02 08:14:23 INFO mapreduce.Job:  map 100% reduce 96%
25/05/02 08:14:25 INFO mapreduce.Job:  map 100% reduce 100%
25/05/02 08:14:25 INFO mapreduce.Job: Job job_1746171991928_0003 completed successfully
25/05/02 08:14:25 INFO mapreduce.Job: Counters: 49
```


Logged in as: dr.who
All Applications
Cluster
About
Nodes
Node Labels
Applications
NEW
NEW_SAVING
SUBMITTED
ACCEPTED
RUNNING
FINISHED
FAILED
KILLED
Scheduler
Tools
Cluster Metrics
Apps Submitted 1
Apps Pending 0
Apps Running 0
Apps Completed 1
Containers Running 0
Memory Used 0 B
Memory Total 16 GB
Memory Reserved 0 B
VCores Used 0
VCores Total 16
VCores Reserved 0
Active Nodes 2
Decommissioned Nodes 0
Lost Nodes 0
Unhealthy Nodes 0
Rebooted Nodes 0
Scheduler Metrics
Scheduler Type: Capacity Scheduler
Scheduling Resource Type: [MEMORY]
Minimum Allocation: <memory:1024, vCores:1>
Maximum Allocation: <memory:8192, vCores:8>
Show 20 entries
Search:
ID: application_1746171991928_0003
User: root
Name: streamjob8088453024948709999.jar
Application Type: MAPREDUCE
Queue: default
StartTime: Fri May 2 09:13:33 +0100 2025
FinishTime: Fri May 2 09:14:24 +0100 2025
State: FINISHED
FinalStatus: SUCCEEDED
Progress
Tracking UI: History
Blacklisted Nodes: N/A
Showing 1 to 1 of 1 entries
First Previous 1 Next Last


Show 20 entries
Search:
ID: application_1746171991928_0003
User: root
Name: streamjob8088453024948709999.jar
Application Type: MAPREDUCE
Queue: default
StartTime: Fri May 2 09:13:33 +0100 2025
FinishTime: Fri May 2 09:14:24 +0100 2025
State: FINISHED
FinalStatus: SUCCEEDED
Progress
Tracking UI: History
Blacklisted Nodes: N/A
Showing 1 to 1 of 1 entries
First Previous 1 Next Last


root@hadoop-master:~# hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar -files mapper.py,reducer.py -mapper "python3 mapper.py" -reducer "python3 reducer.py" -input /user/root/input/ratings.csv -output /user/hadoop/output
packageJobJar: [/tmp/hadoop-unjar4757050382731597l/] [] /tmp/streamjob8088453024948709999.jar tmpDir=null
25/05/02 08:13:32 INFO client.RMProxy: Connecting to ResourceManager at hadoop-master/172.18.0.2:8032
25/05/02 08:13:32 INFO client.RMProxy: Connecting to ResourceManager at hadoop-master/172.18.0.2:8032
25/05/02 08:13:33 INFO mapred.FileInputFormat: Total input paths to process : 1
25/05/02 08:13:33 INFO net.NetworkTopology: Adding a new node: /default-rack/172.18.0.3:50010
25/05/02 08:13:33 INFO net.NetworkTopology: Adding a new node: /default-rack/172.18.0.4:50010
25/05/02 08:13:33 INFO mapreduce.JobSubmitter: number of splits:5
25/05/02 08:13:33 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1746171991928_0003
25/05/02 08:13:33 INFO impl.YarnClientImpl: Submitted application application_1746171991928_0003
25/05/02 08:13:33 INFO mapreduce.Job: The url to track the job: http://hadoop-master:8088/proxy/application_1746171991928_0003/
25/05/02 08:13:33 INFO mapreduce.Job: Running job: job_1746171991928_0003
25/05/02 08:13:38 INFO mapreduce.Job: Job job_1746171991928_0003 running in uber mode : false
25/05/02 08:13:38 INFO mapreduce.Job: map 0% reduce 0%
25/05/02 08:13:47 INFO mapreduce.Job: map 6% reduce 0%
25/05/02 08:13:49 INFO mapreduce.Job: map 31% reduce 0%
25/05/02 08:13:50 INFO mapreduce.Job: map 34% reduce 0%
25/05/02 08:13:52 INFO mapreduce.Job: map 46% reduce 0%
25/05/02 08:13:53 INFO mapreduce.Job: map 48% reduce 0%
25/05/02 08:13:55 INFO mapreduce.Job: map 54% reduce 0%
25/05/02 08:13:56 INFO mapreduce.Job: map 55% reduce 0%
25/05/02 08:13:58 INFO mapreduce.Job: map 63% reduce 0%
25/05/02 08:13:59 INFO mapreduce.Job: map 65% reduce 0%
25/05/02 08:14:01 INFO mapreduce.Job: map 79% reduce 0%
25/05/02 08:14:02 INFO mapreduce.Job: map 94% reduce 0%
25/05/02 08:14:03 INFO mapreduce.Job: map 100% reduce 0%
25/05/02 08:14:11 INFO mapreduce.Job: map 100% reduce 76%
25/05/02 08:14:14 INFO mapreduce.Job: map 100% reduce 81%
25/05/02 08:14:17 INFO mapreduce.Job: map 100% reduce 86%
25/05/02 08:14:20 INFO mapreduce.Job: map 100% reduce 91%
25/05/02 08:14:23 INFO mapreduce.Job: map 100% reduce 96%
25/05/02 08:14:25 INFO mapreduce.Job: map 100% reduce 100%
25/05/02 08:14:25 INFO mapreduce.Job: Job job_1746171991928_0003 completed successfully
25/05/02 08:14:25 INFO mapreduce.Job: Counters: 49


Hadoop
Overview
Datanodes
Snapshot
Startup Progress
Utilities
Browse Directory
/user/hadoop/output
Go!
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
-rw-r--r--
root
supergroup
0 B
02/05/2025 09:14:24
2
128 MB
_SUCCESS
-rw-r--r--
root
supergroup
664.07 KB
02/05/2025 09:14:24
2
128 MB
part-00000
Hadoop, 2015.

```
root@hadoop-master:~# hadoop fs -tail /user/hadoop/output/part-00000
1       2.67
99613   3.37
99615   2.40
99633   3.01
99636   3.61
99638   3.62
99640   3.27
99642   3.31
99644   2.80
99659   2.62
99663   2.38
99665   2.88
99667   3.50
99669   3.05
99671   3.00
99673   3.50
99675   3.48
99677   2.36
99687   2.71
99689   3.21
99691   2.93
99695   3.00
99699   2.25
997     3.53
99706   3.20
99708   3.08
99717   3.50
99721   2.60
99724   3.33
99726   2.67
99728   3.19
99731   2.61
99733   2.59
99735   2.50
99737   2.83
99739   2.08
99741   3.26
99744   2.77
```

Mapreduce 2 :

In the second phase, MapReduce 2, the goal is to enrich the average rating data by joining it with the corresponding movie titles from the `movies.csv` file. This phase uses the output from MapReduce 1 along with `movies.csv` located at `/user/root/input/movies.csv`. The Mapper reads both datasets, associates each average rating with the movie title, and emits key-value pairs where the key is the average rating and the value is the movie title. The Reducer then collects all the results, sorts them in descending order of average rating, and outputs the top 20 highest-rated movies along with their average ratings. The final output is a ranked list in the format: `average_rating movie_title`, which can be used to highlight the most appreciated movies by viewers.

mapper1.py > ...

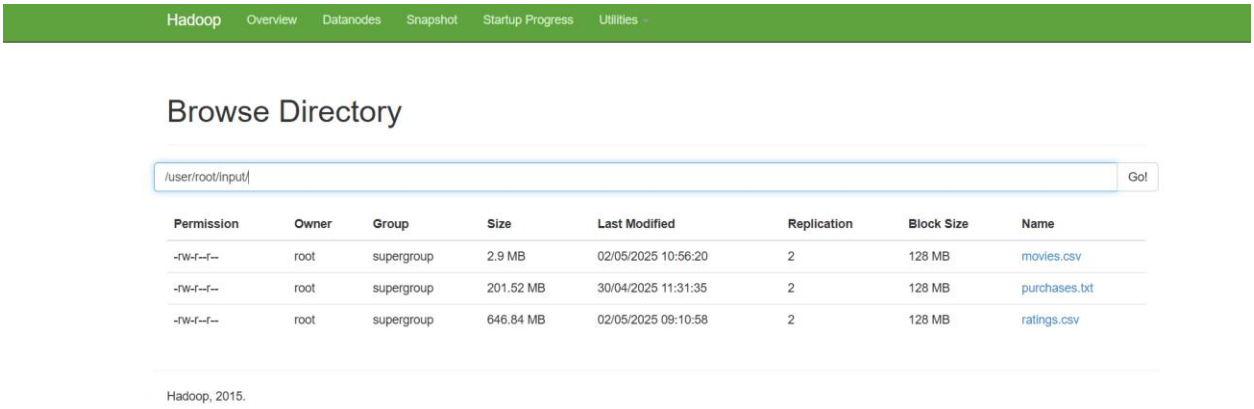
```python
#!/usr/bin/env python3
import sys

for line in sys.stdin:
    line = line.strip()
    if "," in line and line.count(",") >= 2:
        # Likely from movies.csv
        parts = line.split(",", 2)
        if len(parts) >= 2:
            movie_id = parts[0]
            title = parts[1]
            print("{}\tM:{}".format(movie_id, title))
    elif "\t" in line:
        # Likely from ratings average output
        parts = line.split("\t")
        if len(parts) == 2:
            movie_id, avg = parts
            print("{}\tR:{}".format(movie_id, avg))

```

reducer1.py > ...

```python

for line in sys.stdin:
    line = line.strip()
    parts = line.split("\t", 1)
    if len(parts) != 2:
        continue

    movie_id, value = parts

    if current_movie != movie_id:
        if title and avg:
            print("{:.2f}\t{}".format(float(avg), title))
        current_movie = movie_id
        title = None
        avg = None

    if value.startswith("M:"):
        title = value[2:]
    elif value.startswith("R:"):
        avg = value[2:]

# Output last movie
if title and avg:
    print("{:.2f}\t{}".format(float(avg), title))

```

```
C:\Users\Poste>docker cp C:\Users\Poste\Desktop\hdp\mapper1.py myhadoop-master:/root/mapper1.py
Successfully copied 2.56kB to myhadoop-master:/root/mapper1.py

C:\Users\Poste>docker cp C:\Users\Poste\Desktop\hdp\reducer1.py myhadoop-master:/root/reducer1.py
Successfully copied 2.56kB to myhadoop-master:/root/reducer1.py

C:\Users\Poste>docker cp C:\Users\Poste\Desktop\hdp\movies.csv myhadoop-master:/root/movies.csv
Successfully copied 3.04MB to myhadoop-master:/root/movies.csv
```

```
root@hadoop-master:~# hadoop fs -put movies.csv /user/root/input/movies.csv
```




```
root@hadoop-master:~# hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar -files mapper1.py,reducer1.py,/root/movies.csv -input /user/hadoop/output/part-00000 /user/root/input/movies.c
sv  -output /user/hadoop/output_top20 -mapper "python3 mapper1.py"  -reducer "python3 reducer1.py"
packageJobJar: [/tmp/hadoop-unjar8040898478922652318/] [] /tmp/streamjob7538819063419966761.jar tmpDir=null
25/05/02 10:17:15 INFO client.RMProxy: Connecting to ResourceManager at hadoop-master/172.18.0.2:8032
25/05/02 10:17:16 INFO client.RMProxy: Connecting to ResourceManager at hadoop-master/172.18.0.2:8032
25/05/02 10:17:16 INFO mapred.FileInputFormat: Total input paths to process : 2
25/05/02 10:17:16 INFO mapreduce.JobSubmitter: number of splits:3
25/05/02 10:17:16 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1746179619802_0001
25/05/02 10:17:17 INFO impl.YarnClientImpl: Submitted application application_1746179619802_0001
25/05/02 10:17:17 INFO mapreduce.Job: The url to track the job: http://hadoop-master:8088/proxy/application_1746179619802_0001/
25/05/02 10:17:17 INFO mapreduce.Job: Running job: job_1746179619802_0001
25/05/02 10:17:21 INFO mapreduce.Job: Job job_1746179619802_0001 running in uber mode : false
25/05/02 10:17:21 INFO mapreduce.Job:  map 0% reduce 0%
25/05/02 10:17:25 INFO mapreduce.Job:  map 100% reduce 0%
25/05/02 10:17:30 INFO mapreduce.Job:  map 100% reduce 100%
25/05/02 10:17:30 INFO mapreduce.Job: Job job_1746179619802_0001 completed successfully
25/05/02 10:17:30 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=3101547
                FILE: Number of bytes written=6686975
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3722515
                HDFS: Number of bytes written=1739486
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=3
                Launched reduce tasks=1
                Data-local map tasks=3
                Total time spent by all maps in occupied slots (ms)=6422
                Total time spent by all reduces in occupied slots (ms)=1864
                Total time spent by all map tasks (ms)=6422
                Total time spent by all reduce tasks (ms)=1864
                Total vcore-milliseconds taken by all map tasks=6422
                Total vcore-milliseconds taken by all reduce tasks=1864
                Total megabyte-milliseconds taken by all map tasks=6576128
                Total megabyte-milliseconds taken by all reduce tasks=1908736
```

Show 20 entries Search:

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tracking UI	Blacklisted Nodes
application_1746179619802_0003	root	streamjob7479322823409710848.jar	MAPREDUCE	default	Fri May 2 11:25:23 +0100 2025	Fri May 2 11:25:35 +0100 2025	FINISHED	SUCCEEDED		History	N/A
application_1746179619802_0002	root	streamjob6821695598492828465.jar	MAPREDUCE	default	Fri May 2 11:24:19 +0100 2025	Fri May 2 11:24:31 +0100 2025	FINISHED	SUCCEEDED		History	N/A
application_1746179619802_0001	root	streamjob7538819063419966761.jar	MAPREDUCE	default	Fri May 2 11:17:16 +0100 2025	Fri May 2 11:17:28 +0100 2025	FINISHED	SUCCEEDED		History	N/A

Showing 1 to 3 of 3 entries First Previous 1 Next Last

```
25/05/02 10:17:30 INFO streaming.StreamJob: Output directory: /user/hadoop/output_top20
root@hadoop-master:~# hadoop fs -tail /user/hadoop/output_top20
tail: `/user/hadoop/output_top20': Is a directory
root@hadoop-master:~# hadoop fs -tail /user/hadoop/output_top20/part-00000
77)
2.36    Codependent Lesbian Space Alien Seeks Same (2011)
2.93    "World's Greatest Athlete
2.33    Cargo (2011)
3.50    Diana Vreeland: The Eye Has to Travel (2011)
3.50    Sangre de mi sangre (Padre Nuestro) (2007)
3.14    Milarepa (2006)
3.28    2 Days in the Valley (1996)
1.75    O Panishyros Megistanas Ton Ninja (2008)
2.17    Morgan Pålsson - världsreporter (2008)
3.10    Renoir (2012)
3.07    "Last Stand
3.13    Mama (2013)
3.17    "Candidate
3.58    Upstream Color (2013)
3.00    Fast Life (1932)
2.83    Careless Love (2012)
3.67    Special When Lit  (2009)
2.50    "Year of the Tiger
4.00    Preston Sturges: The Rise and Fall of an American Dreamer (1990)
2.38    Double Indemnity (1973)
3.60    London After Midnight (2002)
2.99    Broken City (2013)
4.25    Nursery University (2008)
3.62    "Counterfeit Coin
3.70    Nameless Gangster (Bumchoiwaui junjaeng) (2012)
3.12    Ginger & Rosa (2012)
3.67    First Shot (2002)
3.52    Marina Abramovic: The Artist Is Present (2012)
3.94    Bonsái (2011)
3.28    Shadow Dancer (2012)
3.52    It's a Disaster (2012)
3.60    Texas Across the River (1966)
```

Overall, this two-stage MapReduce project demonstrates how distributed processing can be used to perform data aggregation and enrichment on large datasets. It provides a scalable approach to computing movie recommendations based on user ratings and movie metadata.

Mapreduce 3: Highest Rated Movies for Each Genre

mapperg.py > ...

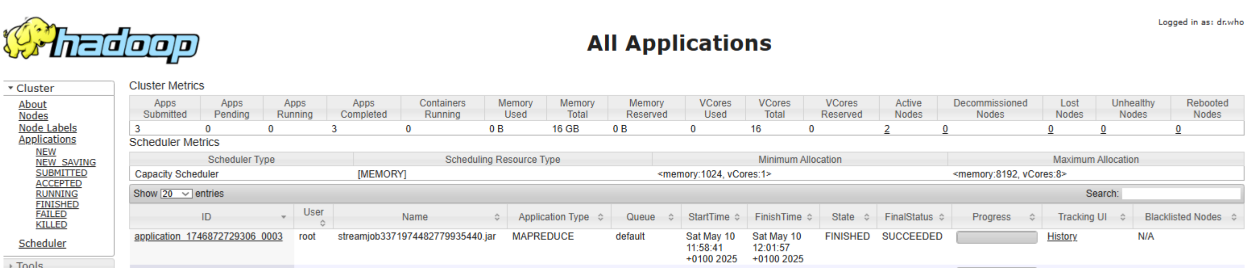
```python
#!/usr/bin/env python3
import sys

# Read each line from standard input
for line in sys.stdin:
    fields = line.strip().split(",")

    # Skip the header
    if fields[0] == "userId":
        continue

    # Extract movieId and rating
    _, movie_id, rating, _ = fields

    print(f"{movie_id}\t{rating}")

```

reducerg.py > ...

```python
#!/usr/bin/env python3
import sys
from collections import defaultdict

# Load the movie genres from the file
movies = {}
with open('/data/movies.csv', 'r') as file:
    next(file)  # skip header
    for line in file:
        movie_id, title, genres = line.strip().split(",", 2)
        movies[movie_id] = (title, genres.split("|"))

# Initialize containers
movie_ratings = defaultdict(list)

# Read input from mapper
for line in sys.stdin:
    movie_id, rating = line.strip().split("\t")
    if movie_id in movies:
        movie_ratings[movie_id].append(float(rating))

# Calculate average and group by genre
genre_top_movies = defaultdict(list)

for movie_id, ratings in movie_ratings.items():
    title, genres = movies[movie_id]
    average = sum(ratings) / len(ratings)
    for genre in genres:
        genre_top_movies[genre].append((average, title))
```

```
root@hadoop-master:~# hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar -files mapperg.py,reducerg.py,/root/movies.csv
-mapper "python3 mapperg.py" -reducer "python3 reducerg.py" -input /user/root/input/ratings.csv -output /user/hadoop/output/top_movie_per_genre
packageJobJar: [/tmp/hadoop-unjar4381728813468423454/] [] /tmp/streamjob3371974482779935440.jar tmpDir=null
25/05/10 10:58:40 INFO client.RMProxy: Connecting to ResourceManager at hadoop-master/172.18.0.4:8032
25/05/10 10:58:40 INFO client.RMProxy: Connecting to ResourceManager at hadoop-master/172.18.0.4:8032
25/05/10 10:58:41 INFO mapred.FileInputFormat: Total input paths to process : 1
25/05/10 10:58:41 INFO net.NetworkTopology: Adding a new node: /default-rack/172.18.0.2:50010
25/05/10 10:58:41 INFO net.NetworkTopology: Adding a new node: /default-rack/172.18.0.3:50010
25/05/10 10:58:41 INFO mapreduce.JobSubmitter: number of splits:5
25/05/10 10:58:41 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1746872729306_0003
25/05/10 10:58:41 INFO impl.YarnClientImpl: Submitted application application_1746872729306_0003
25/05/10 10:58:41 INFO mapreduce.Job: The url to track the job: http://hadoop-master:8088/proxy/application_1746872729306_0003/
25/05/10 10:58:41 INFO mapreduce.Job: Running job: job_1746872729306_0003
25/05/10 10:58:46 INFO mapreduce.Job: Job job_1746872729306_0003 running in uber mode : false
25/05/10 10:58:46 INFO mapreduce.Job:  map 0% reduce 0%
25/05/10 10:58:56 INFO mapreduce.Job:  map 14% reduce 0%
25/05/10 10:58:59 INFO mapreduce.Job:  map 20% reduce 0%
25/05/10 10:59:02 INFO mapreduce.Job:  map 27% reduce 0%
25/05/10 10:59:05 INFO mapreduce.Job:  map 33% reduce 0%
25/05/10 10:59:08 INFO mapreduce.Job:  map 39% reduce 0%
25/05/10 10:59:11 INFO mapreduce.Job:  map 46% reduce 0%
25/05/10 10:59:14 INFO mapreduce.Job:  map 52% reduce 0%
25/05/10 10:59:17 INFO mapreduce.Job:  map 59% reduce 0%
25/05/10 10:59:20 INFO mapreduce.Job:  map 65% reduce 0%
25/05/10 10:59:23 INFO mapreduce.Job:  map 75% reduce 0%
25/05/10 10:59:26 INFO mapreduce.Job:  map 86% reduce 0%
25/05/10 10:59:28 INFO mapreduce.Job:  map 90% reduce 0%
25/05/10 10:59:29 INFO mapreduce.Job:  map 98% reduce 0%
25/05/10 10:59:30 INFO mapreduce.Job:  map 100% reduce 0%
25/05/10 10:59:38 INFO mapreduce.Job:  map 100% reduce 67%
25/05/10 10:59:41 INFO mapreduce.Job:  map 100% reduce 70%
25/05/10 10:59:44 INFO mapreduce.Job:  map 100% reduce 71%
25/05/10 10:59:50 INFO mapreduce.Job:  map 100% reduce 74%
25/05/10 10:59:53 INFO mapreduce.Job:  map 100% reduce 77%
25/05/10 10:59:56 INFO mapreduce.Job:  map 100% reduce 79%
25/05/10 11:00:05 INFO mapreduce.Job:  map 100% reduce 82%
25/05/10 11:00:08 INFO mapreduce.Job:  map 100% reduce 84%
25/05/10 11:00:17 INFO mapreduce.Job:  map 100% reduce 88%
25/05/10 11:00:20 INFO mapreduce.Job:  map 100% reduce 91%
25/05/10 11:00:23 INFO mapreduce.Job:  map 100% reduce 92%
25/05/10 11:00:42 INFO mapreduce.Job:  map 100% reduce 94%
25/05/10 11:00:52 INFO mapreduce.Job:  map 100% reduce 96%
25/05/10 11:00:58 INFO mapreduce.Job:  map 100% reduce 99%
25/05/10 11:01:01 INFO mapreduce.Job:  map 100% reduce 100%
25/05/10 11:01:59 INFO mapreduce.Job: Job job_1746872729306_0003 completed successfully
25/05/10 11:02:00 INFO mapreduce.Job: Counters: 49
```

```
root@hadoop-master:~# hadoop fs -cat /user/hadoop/output/top_movie_per_genre/part-00000 | head -n 50
The (2007)",Adventure	56171	5.00
The (2007)",Adventure	55269	5.00
The (2007)",Adventure	55269	5.00
The (2007)",Adventure	55269	5.00
The (2007)",Adventure	55269	5.00
The (2007)",Adventure	55269	5.00
The (2007)",Adventure	56171	5.00
The (2007)",Adventure	56171	5.00
The (2007)",Adventure	55269	5.00
The (2007)",Adventure	55269	5.00
The (2007)",Adventure	56171	5.00
The (2007)",Adventure	55269	5.00
The (2007)",Adventure	56171	5.00
The (2007)",Adventure	55269	5.00
The (2007)",Adventure	55269	5.00
The (2007)",Adventure	55269	5.00
The (2007)",Adventure	55269	5.00
The (2007)",Adventure	55269	5.00
The (2007)",Adventure	55269	5.00
The (2007)",Adventure	55269	5.00
The (1980)",Action	1220	5.00
The (1980)",Action	1220	5.00
The (1980)",Action	1220	5.00
The (1980)",Action	1220	5.00
The (1980)",Action	1220	5.00
The (1980)",Action	1220	5.00
The (1980)",Action	1220	5.00
The (1980)",Action	1220	5.00
The (1980)",Action	1220	5.00
The (1980)",Action	1220	5.00
The (1980)",Action	1220	5.00
The (1980)",Action	1220	5.00
The (1980)",Action	1220	5.00
The (1980)",Action	1220	5.00
The (1980)",Action	1220	5.00
The (1980)",Action	1220	5.00
The (1980)",Action	1220	5.00
The (1980)",Action	1220	5.00
The (1980)",Action	1220	5.00
The (1980)",Action	1220	5.00
The (Le amiche) (1955)",Drama	81413	5.00
The (Le amiche) (1955)",Drama	81413	4.50
The (Le amiche) (1955)",Drama	81413	4.50
The (Le amiche) (1955)",Drama	81413	4.50
The (Le amiche) (1955)",Drama	81413	4.00
The (Le amiche) (1955)",Drama	81413	4.00
The (Le amiche) (1955)",Drama	81413	4.00
The (Le amiche) (1955)",Drama	81413	4.00
```



In this MapReduce job, the mapper processed each rating entry, associating it with the movie's name and genre. This allowed the reducer to group all ratings for the same movie and genre combination. The output reflected each rating received, providing a detailed view of every instance where a rating was registered. This level of granularity ensured that all interactions were captured, but it displayed multiple lines for movies with multiple ratings in the same genre.

Mapreduce 4:The average rating for each film per genre

The mapper still emitted the movie's name and genre as the key, along with its rating. However, in the reducer step, the logic was enhanced to **calculate the average rating** for each movie within each genre. This aggregation step provided a consolidated view, where each movie appeared just once per genre with its average rating, making the output clearer and more insightful.

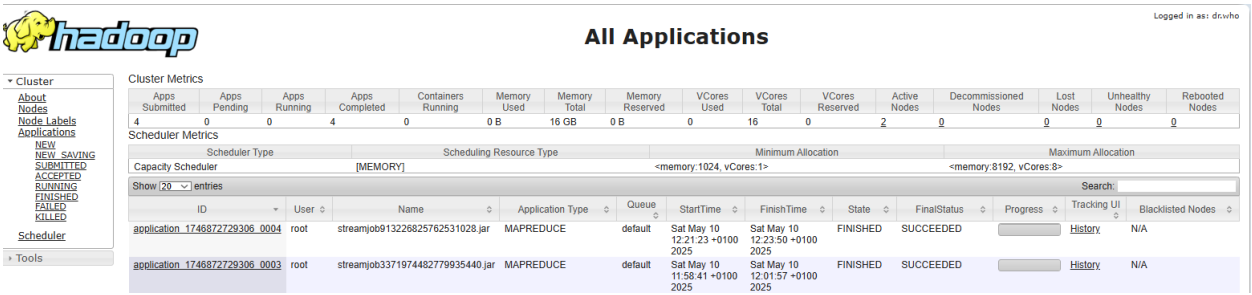
mapperg1.py > ...

```
# Load the movies data into a dictionary
movies = {}
with open('movies.csv', 'r', encoding='utf-8') as f:
    next(f)  # Skip header
    for line in f:
        fields = line.strip().split(',', 2)
        movie_id, title, genres = fields[0], fields[1], fields[2]
        movies[movie_id] = (title, genres)

# Process ratings input
for line in sys.stdin:
    try:
        fields = line.strip().split(',')
        if len(fields) < 4:
            continue
        movie_id, rating = fields[1], float(fields[2])

        if movie_id in movies:
            title, genres = movies[movie_id]
            for genre in genres.split('|'):
                # Output format: genre \t movie_name \t rating
                print("{}\t{}\t{}".format(genre, title, rating))
    except Exception as e:
        continue

```

reducerg1.py > ...

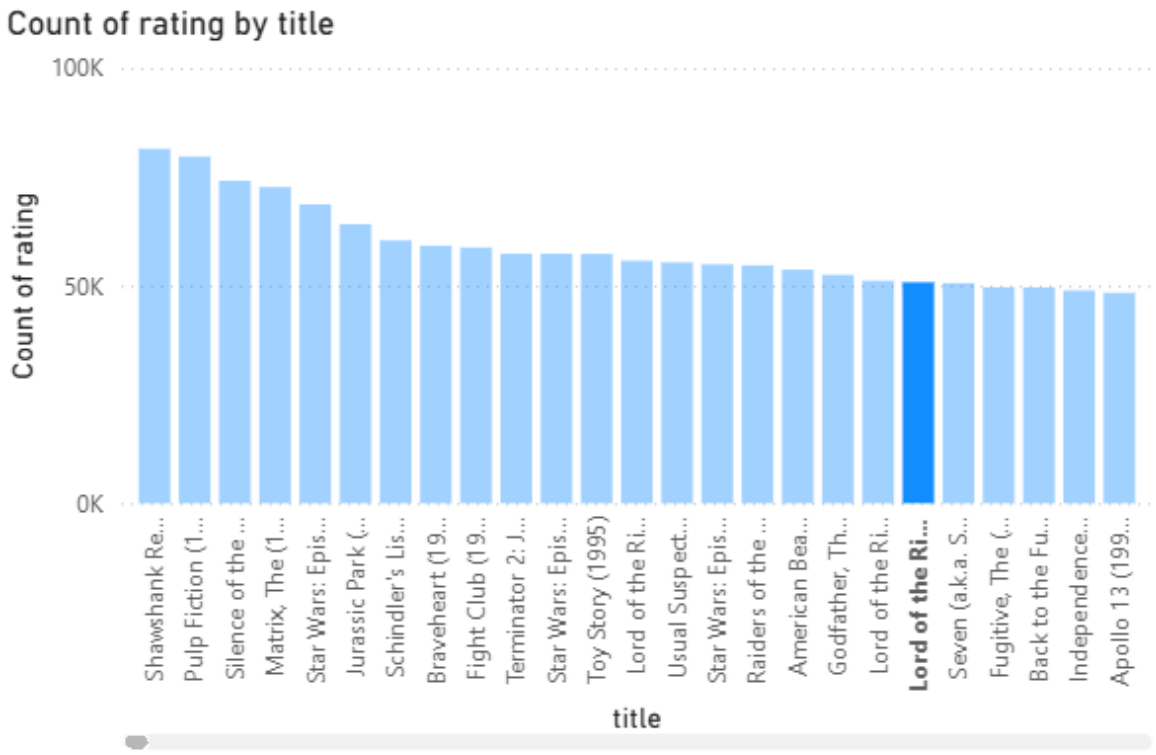
```python
#!/usr/bin/env python3
import sys
from collections import defaultdict

# Initialize storage
genre_movies = defaultdict(lambda: defaultdict(list))

# Read input from Mapper
for line in sys.stdin:
    try:
        genre, title, rating = line.strip().split('\t')
        rating = float(rating)
        genre_movies[genre][title].append(rating)
    except ValueError:
        continue

# Process and sort movies by genre
for genre, movies in genre_movies.items():
    # Calculate the average rating for each movie
    sorted_movies = sorted(
        [(title, sum(ratings) / len(ratings)) for title, ratings in movies.items()],
        key=lambda x: -x[1]
    )

    # Output the top 20 movies in each genre
    for title, average in sorted_movies[:20]:
        print("{}\t{}\t{:.2f}".format(genre, title, average))

```

```
put /user/root/input/ratings.csv -output /user/hadoop/output/top_movies_average_per_genre
packageJobJar: [/tmp/hadoop-unjar7545141381103462411/] [] /tmp/streamjob9132268257625310 28.jar tmpDir=null
25/05/10 11:21:22 INFO client.RMProxy: Connecting to ResourceManager at hadoop-master/172.18.0.4:8032
25/05/10 11:21:22 INFO client.RMProxy: Connecting to ResourceManager at hadoop-master/172.18.0.4:8032
25/05/10 11:21:23 INFO mapred.FileInputFormat: Total input paths to process : 1
25/05/10 11:21:23 INFO net.NetworkTopology: Adding a new node: /default-rack/172.18.0.3:50010
25/05/10 11:21:23 INFO net.NetworkTopology: Adding a new node: /default-rack/172.18.0.2:50010
25/05/10 11:21:23 INFO mapreduce.JobSubmitter: number of splits:5
25/05/10 11:21:23 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1746872729306_0004
25/05/10 11:21:23 INFO impl.YarnClientImpl: Submitted application application_1746872729306_0004
25/05/10 11:21:23 INFO mapreduce.Job: The url to track the job: http://hadoop-master:8088/proxy/application_1746872729306_0004/
25/05/10 11:21:23 INFO mapreduce.Job: Running job: job_1746872729306_0004
25/05/10 11:21:28 INFO mapreduce.Job: Job job_1746872729306_0004 running in uber mode : false
25/05/10 11:21:28 INFO mapreduce.Job:  map 0% reduce 0%
25/05/10 11:21:38 INFO mapreduce.Job:  map 9% reduce 0%
25/05/10 11:21:41 INFO mapreduce.Job:  map 14% reduce 0%
25/05/10 11:21:44 INFO mapreduce.Job:  map 19% reduce 0%
25/05/10 11:21:47 INFO mapreduce.Job:  map 24% reduce 0%
25/05/10 11:21:50 INFO mapreduce.Job:  map 29% reduce 0%
25/05/10 11:21:53 INFO mapreduce.Job:  map 34% reduce 0%
25/05/10 11:21:56 INFO mapreduce.Job:  map 39% reduce 0%
25/05/10 11:21:59 INFO mapreduce.Job:  map 44% reduce 0%
25/05/10 11:22:02 INFO mapreduce.Job:  map 49% reduce 0%
25/05/10 11:22:05 INFO mapreduce.Job:  map 54% reduce 0%
25/05/10 11:22:08 INFO mapreduce.Job:  map 58% reduce 0%
25/05/10 11:22:11 INFO mapreduce.Job:  map 63% reduce 0%
25/05/10 11:22:14 INFO mapreduce.Job:  map 68% reduce 0%
25/05/10 11:22:17 INFO mapreduce.Job:  map 79% reduce 0%
25/05/10 11:22:20 INFO mapreduce.Job:  map 89% reduce 0%
25/05/10 11:22:22 INFO mapreduce.Job:  map 91% reduce 0%
25/05/10 11:22:23 INFO mapreduce.Job:  map 98% reduce 0%
25/05/10 11:22:24 INFO mapreduce.Job:  map 100% reduce 0%
25/05/10 11:22:33 INFO mapreduce.Job:  map 100% reduce 27%
25/05/10 11:22:36 INFO mapreduce.Job:  map 100% reduce 67%
25/05/10 11:22:39 INFO mapreduce.Job:  map 100% reduce 69%
25/05/10 11:22:42 INFO mapreduce.Job:  map 100% reduce 70%
25/05/10 11:22:48 INFO mapreduce.Job:  map 100% reduce 73%
25/05/10 11:22:54 INFO mapreduce.Job:  map 100% reduce 77%
25/05/10 11:22:57 INFO mapreduce.Job:  map 100% reduce 78%
25/05/10 11:23:06 INFO mapreduce.Job:  map 100% reduce 82%
25/05/10 11:23:09 INFO mapreduce.Job:  map 100% reduce 83%
25/05/10 11:23:18 INFO mapreduce.Job:  map 100% reduce 88%
25/05/10 11:23:21 INFO mapreduce.Job:  map 100% reduce 89%
25/05/10 11:23:24 INFO mapreduce.Job:  map 100% reduce 91%
25/05/10 11:23:27 INFO mapreduce.Job:  map 100% reduce 92%
25/05/10 11:23:33 INFO mapreduce.Job:  map 100% reduce 94%
25/05/10 11:23:36 INFO mapreduce.Job:  map 100% reduce 96%
25/05/10 11:23:42 INFO mapreduce.Job:  map 100% reduce 99%
25/05/10 11:23:45 INFO mapreduce.Job:  map 100% reduce 100%
25/05/10 11:23:51 INFO mapreduce.Job: Job job_1746872729306_0004 completed successfully
```




```
Action      "Getaway        3.61
Action      "Mechanic       3.43
Action      "Poseidon Adventure     3.31
Action      "Big Bird Cage  3.18
Action      "Groundstar Conspiracy  2.92
Action      "Unholy Rollers 2.00
Comedy      "Happiest Days of Your Life     3.39
Comedy      "Yellow Cab Man 3.00
Comedy      "West Point Story       2.92
Comedy      "Big Hangover   2.25
,Drama      "Goodbye Bafana (Color of Freedom       3.57
```

title	Count of rating
"Great Performances" Cats (1998)	10
$5 a Day (2008)	9
$9.99 (2008)	10
(500) Days of Summer (2009)	10
(Untitled) (2009)	10
*batteries not included (1987)	10
...All the Marbles (1981)	10
...And God Spoke (1993)	10
...And Justice for All (1979)	10
.45 (2006)	9
[REC] (2007)	10
[REC] 4: Apocalypse (2014)	10
[REC]² (2009)	10
[REC]³ 3 Génesis (2012)	10
¡Three Amigos! (1986)	10
+1 (2013)	10
1 (2014)	9
10 (1979)	10
10 Cloverfield Lane (2016)	10
10 Items or Less (2006)	10
10 Rillington Place (1971)	10
10 Things I Hate About You (1999)	10
10 to Midnight (1983)	10
Total	**10**



title	rating
Lord of the Rings: The Return of the King, The (2003)	0.5
Lord of the Rings: The Return of the King, The (2003)	1.0
Lord of the Rings: The Return of the King, The (2003)	1.5
Lord of the Rings: The Return of the King, The (2003)	2.0
Lord of the Rings: The Return of the King, The (2003)	2.5
Lord of the Rings: The Return of the King, The (2003)	3.0
Lord of the Rings: The Return of the King, The (2003)	3.5
Lord of the Rings: The Return of the King, The (2003)	4.0
Lord of the Rings: The Return of the King, The (2003)	4.5
Lord of the Rings: The Return of the King, The (2003)	5.0