Student

Vasseur Pierre-Adrien - M2 Apprenticeship student

Requirements

To run this project, you need Docker & Docker-compose.

Installation

Make sure to have the following files and folders:

- `hadoop.env` > environment variables for the hadoop cluster
- `docker-compose.yml` > docker-compose file to run the hadoop cluster
- `data/` > folder containing the following files :
  - `movies.csv`
  - `ratings.csv`
  - `Movie-1.0-SNAPSHOT.jar`
- `Movie/` > folder containing the source code of the project

Then you can run the following commands :

- `docker compose build --no-cache`
- `docker compose up`

Enter the `namenode` container and run the following commands :

- `hdfs dfs -mkdir /input`
- `hdfs dfs -put ./hadoop/labs/movies.csv /input`
- `hdfs dfs -put ./hadoop/labs/ratings.csv /input`

The jar executable is already packaged in the `data/` folder. This folder is mounted at `./hadoop/labs` in the container.

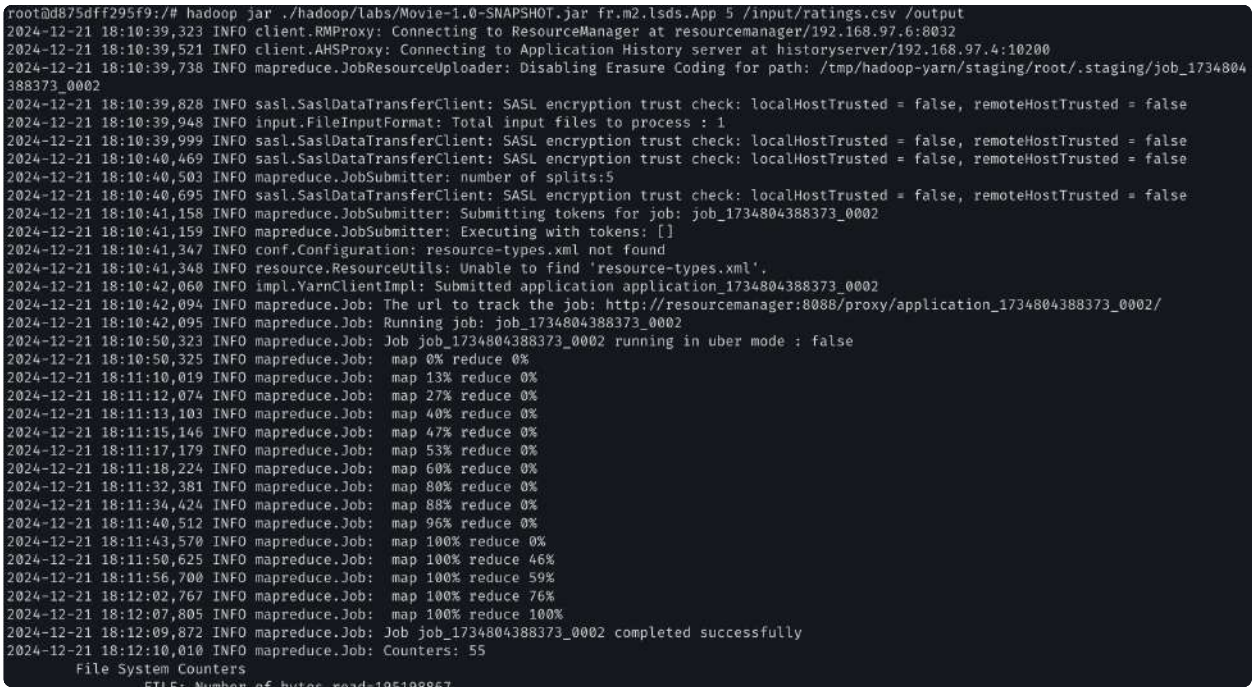
Question 1

> Find the highest rated movieID per user.

Run the following commands :

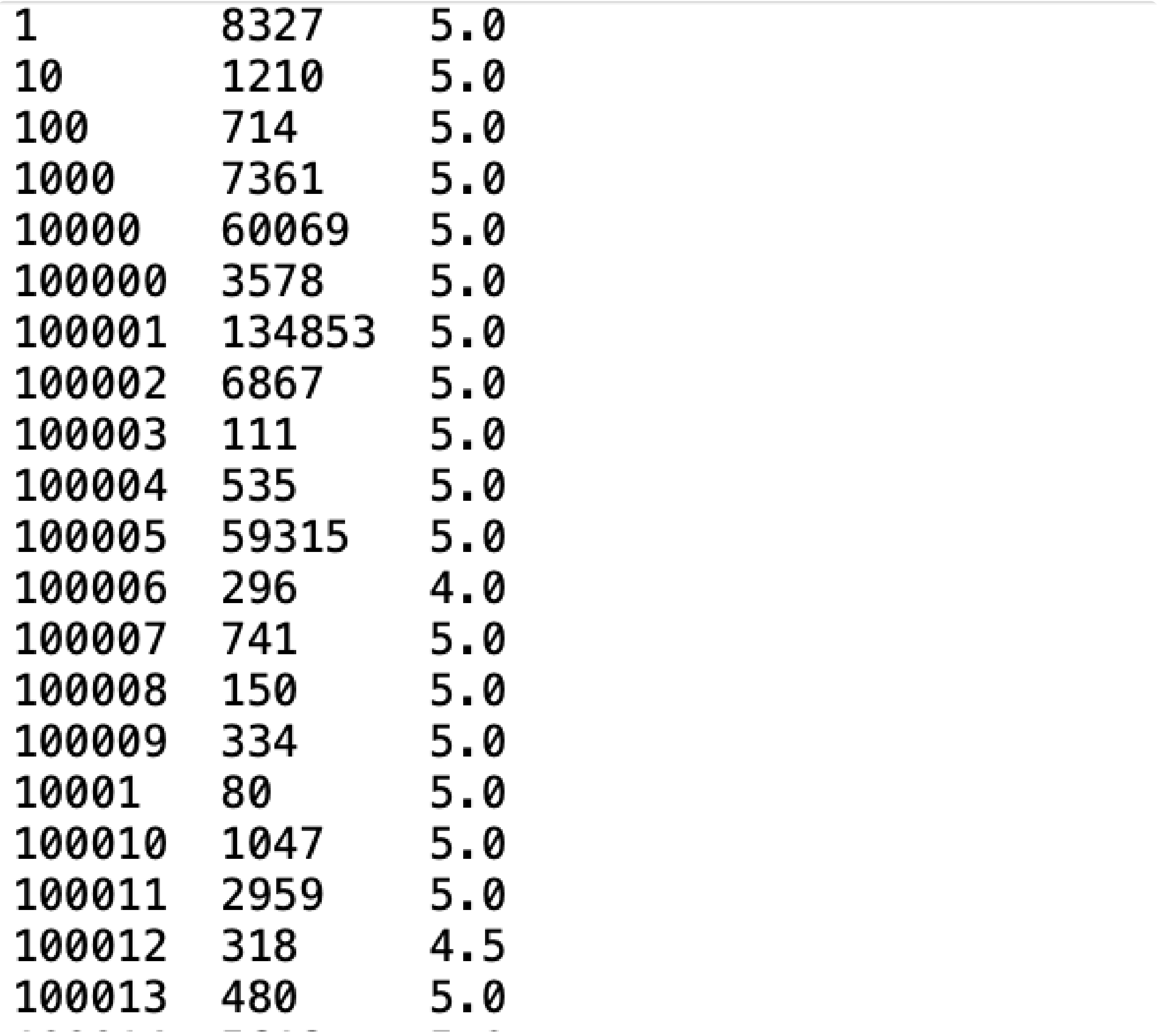
- `hadoop jar ./hadoop/labs/Movie-1.0-SNAPSHOT.jar fr.m2.lsds.App 5 /input/ratings.csv /output`

You should see something like this in the terminal :



```
root@d875dff295f9:/# hadoop jar ./hadoop/labs/Movie-1.0-SNAPSHOT.jar fr.m2.lsds.App 5 /input/ratings.csv /output
2024-12-21 18:10:39,323 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/192.168.97.6:8032
2024-12-21 18:10:39,521 INFO client.AHSProxy: Connecting to Application History server at historyserver/192.168.97.4:10200
2024-12-21 18:10:39,738 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1734804
388373_0002
2024-12-21 18:10:39,828 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-12-21 18:10:39,948 INFO input.FileInputFormat: Total input files to process : 1
2024-12-21 18:10:39,999 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-12-21 18:10:40,469 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-12-21 18:10:40,503 INFO mapreduce.JobSubmitter: number of splits:5
2024-12-21 18:10:40,695 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-12-21 18:10:41,158 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1734804388373_0002
2024-12-21 18:10:41,159 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-12-21 18:10:41,347 INFO conf.Configuration: resource-types.xml not found
2024-12-21 18:10:41,348 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-12-21 18:10:42,060 INFO impl.YarnClientImpl: Submitted application application_1734804388373_0002
2024-12-21 18:10:42,094 INFO mapreduce.Job: The url to track the job: http://resourcemanager:8088/proxy/application_1734804388373_0002/
2024-12-21 18:10:42,095 INFO mapreduce.Job: Running job: job_1734804388373_0002
2024-12-21 18:10:50,323 INFO mapreduce.Job: Job job_1734804388373_0002 running in uber mode : false
2024-12-21 18:10:50,325 INFO mapreduce.Job:  map 0% reduce 0%
2024-12-21 18:11:10,019 INFO mapreduce.Job:  map 13% reduce 0%
2024-12-21 18:11:12,074 INFO mapreduce.Job:  map 27% reduce 0%
2024-12-21 18:11:13,103 INFO mapreduce.Job:  map 40% reduce 0%
2024-12-21 18:11:15,146 INFO mapreduce.Job:  map 47% reduce 0%
2024-12-21 18:11:17,179 INFO mapreduce.Job:  map 53% reduce 0%
2024-12-21 18:11:18,224 INFO mapreduce.Job:  map 60% reduce 0%
2024-12-21 18:11:32,381 INFO mapreduce.Job:  map 80% reduce 0%
2024-12-21 18:11:34,424 INFO mapreduce.Job:  map 88% reduce 0%
2024-12-21 18:11:40,512 INFO mapreduce.Job:  map 96% reduce 0%
2024-12-21 18:11:43,570 INFO mapreduce.Job:  map 100% reduce 0%
2024-12-21 18:11:50,625 INFO mapreduce.Job:  map 100% reduce 46%
2024-12-21 18:11:56,700 INFO mapreduce.Job:  map 100% reduce 59%
2024-12-21 18:12:02,767 INFO mapreduce.Job:  map 100% reduce 76%
2024-12-21 18:12:07,805 INFO mapreduce.Job:  map 100% reduce 100%
2024-12-21 18:12:09,872 INFO mapreduce.Job: Job job_1734804388373_0002 completed successfully
2024-12-21 18:12:10,010 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=195198867
```

And have something like this in the `output` folder :



```
1       8327    5.0
10      1210    5.0
100     714     5.0
1000    7361    5.0
10000   60069   5.0
100000  3578    5.0
100001  134853  5.0
100002  6867    5.0
100003  111     5.0
100004  535     5.0
100005  59315   5.0
100006  296     4.0
100007  741     5.0
100008  150     5.0
100009  334     5.0
10001   80      5.0
100010  1047    5.0
100011  2959    5.0
100012  318     4.5
100013  480     5.0
```

File results are available in the `Movie/results/highestRatedMovieByUserId.txt` file. Code is available in the `Movie/src/main/java/fr/m2/lsds/exercise5` folder.

Question 2

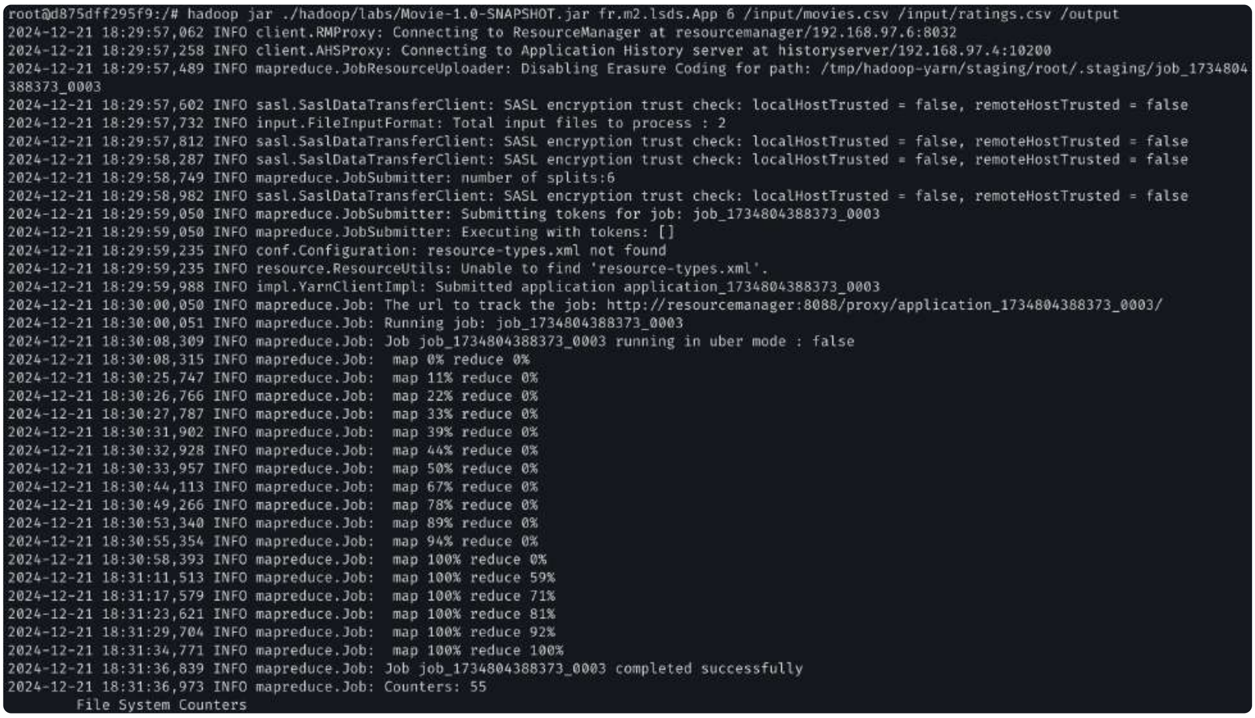
> Find the name of the highest rated movie per userID.

Make sure to delete the `output` before running the next commands otherwise you will have an error.

Run the following commands :

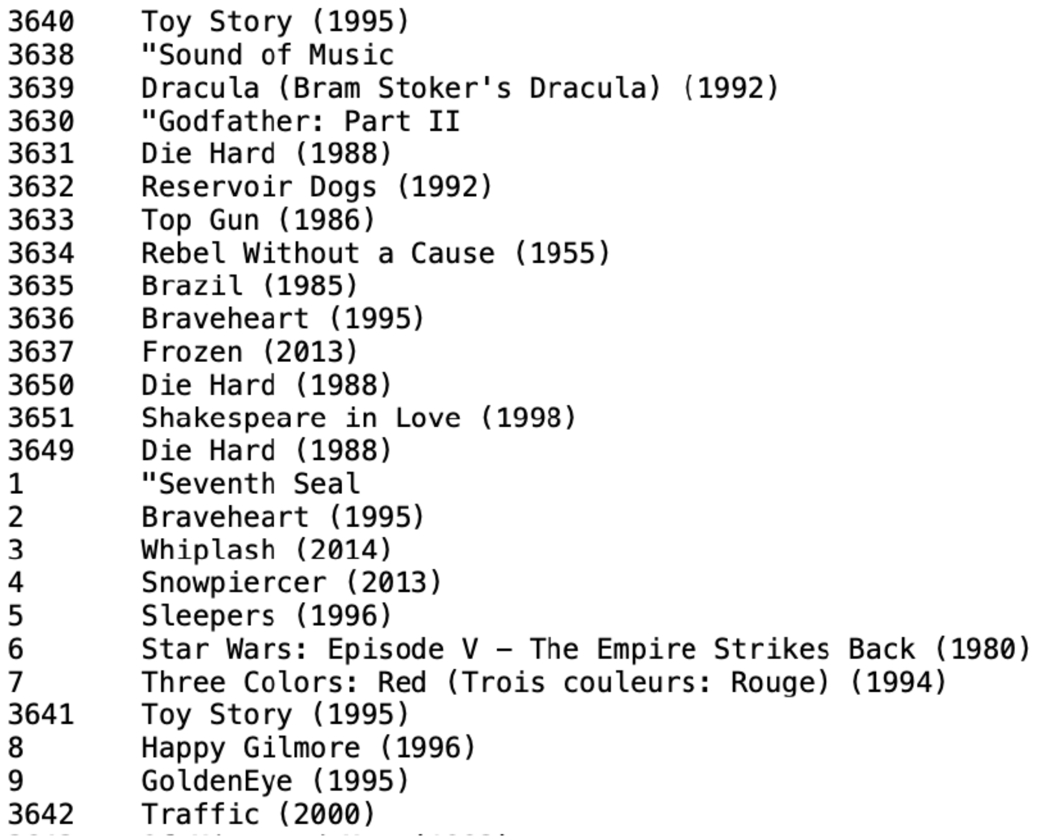
- `hadoop jar ./hadoop/labs/Movie-1.0-SNAPSHOT.jar fr.m2.lsds.App 6 /input/movies.csv /input/ratings.csv /output`

You should see something like this in the terminal :



```
root@d875dff295f9:/# hadoop jar ./hadoop/labs/Movie-1.0-SNAPSHOT.jar fr.m2.lsds.App 6 /input/movies.csv /input/ratings.csv /output
2024-12-21 18:29:57,062 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/192.168.97.6:8032
2024-12-21 18:29:57,258 INFO client.AHSProxy: Connecting to Application History server at historyserver/192.168.97.4:10200
2024-12-21 18:29:57,489 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1734804
388373_0003
2024-12-21 18:29:57,602 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-12-21 18:29:57,732 INFO input.FileInputFormat: Total input files to process : 2
2024-12-21 18:29:57,812 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-12-21 18:29:58,287 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-12-21 18:29:58,749 INFO mapreduce.JobSubmitter: number of splits:6
2024-12-21 18:29:58,982 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-12-21 18:29:59,050 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1734804388373_0003
2024-12-21 18:29:59,050 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-12-21 18:29:59,235 INFO conf.Configuration: resource-types.xml not found
2024-12-21 18:29:59,235 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-12-21 18:29:59,988 INFO impl.YarnClientImpl: Submitted application application_1734804388373_0003
2024-12-21 18:30:00,050 INFO mapreduce.Job: The url to track the job: http://resourcemanager:8088/proxy/application_1734804388373_0003/
2024-12-21 18:30:00,051 INFO mapreduce.Job: Running job: job_1734804388373_0003
2024-12-21 18:30:08,309 INFO mapreduce.Job: Job job_1734804388373_0003 running in uber mode : false
2024-12-21 18:30:08,315 INFO mapreduce.Job:  map 0% reduce 0%
2024-12-21 18:30:25,747 INFO mapreduce.Job:  map 11% reduce 0%
2024-12-21 18:30:26,766 INFO mapreduce.Job:  map 22% reduce 0%
2024-12-21 18:30:27,787 INFO mapreduce.Job:  map 33% reduce 0%
2024-12-21 18:30:31,902 INFO mapreduce.Job:  map 39% reduce 0%
2024-12-21 18:30:32,928 INFO mapreduce.Job:  map 44% reduce 0%
2024-12-21 18:30:33,957 INFO mapreduce.Job:  map 50% reduce 0%
2024-12-21 18:30:44,113 INFO mapreduce.Job:  map 67% reduce 0%
2024-12-21 18:30:49,266 INFO mapreduce.Job:  map 78% reduce 0%
2024-12-21 18:30:53,340 INFO mapreduce.Job:  map 89% reduce 0%
2024-12-21 18:30:55,354 INFO mapreduce.Job:  map 94% reduce 0%
2024-12-21 18:30:58,393 INFO mapreduce.Job:  map 100% reduce 0%
2024-12-21 18:31:11,513 INFO mapreduce.Job:  map 100% reduce 59%
2024-12-21 18:31:17,579 INFO mapreduce.Job:  map 100% reduce 71%
2024-12-21 18:31:23,621 INFO mapreduce.Job:  map 100% reduce 81%
2024-12-21 18:31:29,704 INFO mapreduce.Job:  map 100% reduce 92%
2024-12-21 18:31:34,771 INFO mapreduce.Job:  map 100% reduce 100%
2024-12-21 18:31:36,839 INFO mapreduce.Job: Job job_1734804388373_0003 completed successfully
2024-12-21 18:31:36,973 INFO mapreduce.Job: Counters: 55
        File System Counters
```

And have something like this in the `output` folder :



```
3640    Toy Story (1995)
3638    "Sound of Music
3639    Dracula (Bram Stoker's Dracula) (1992)
3630    "Godfather: Part II
3631    Die Hard (1988)
3632    Reservoir Dogs (1992)
3633    Top Gun (1986)
3634    Rebel Without a Cause (1955)
3635    Brazil (1985)
3636    Braveheart (1995)
3637    Frozen (2013)
3650    Die Hard (1988)
3651    Shakespeare in Love (1998)
3649    Die Hard (1988)
1       "Seventh Seal
2       Braveheart (1995)
3       Whiplash (2014)
4       Snowpiercer (2013)
5       Sleepers (1996)
6       Star Wars: Episode V - The Empire Strikes Back (1980)
7       Three Colors: Red (Trois couleurs: Rouge) (1994)
3641    Toy Story (1995)
8       Happy Gilmore (1996)
9       GoldenEye (1995)
3642    Traffic (2000)
```

File results are available in the `Movie/results/highestRatedMovieNameByUserId.txt` file.
Code is available in the `Movie/src/main/java/fr/m2/lsds/exercise6` folder.

Question 3

> Count how many users liked this movie, and group them by like-count.

Make sure to delete the `output` before running the next commands otherwise you will have an error.

For this question, you need the result of the previous question, so make sure to save it and upload it to the `input` folder.

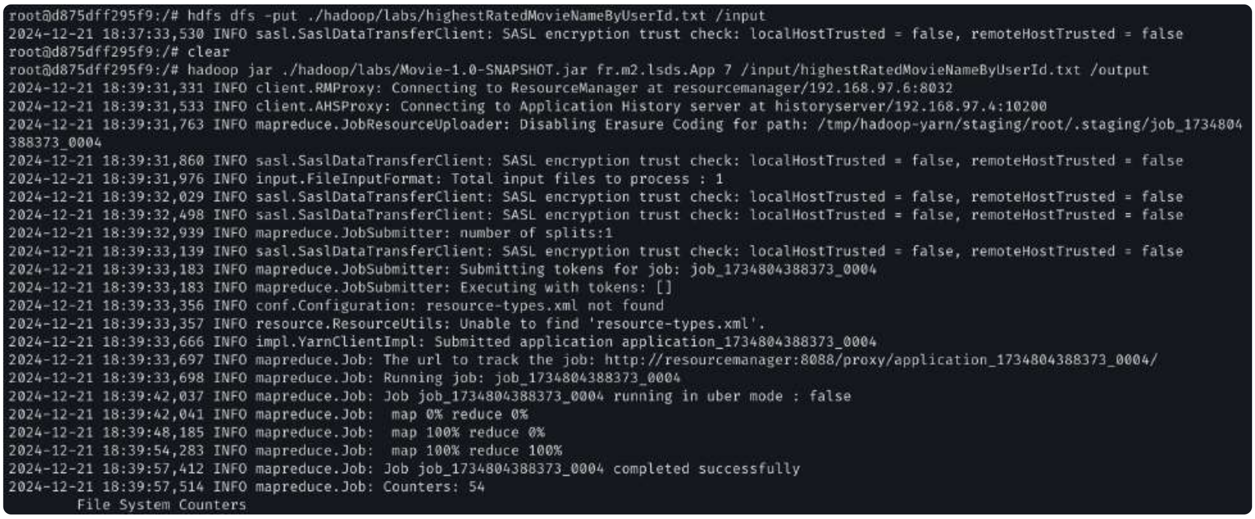
- `hdfs dfs -put ./hadoop/labs/highestRatedMovieNameByUserId.txt /input`

Then run the following commands :

- `hadoop jar ./hadoop/labs/Movie-1.0-SNAPSHOT.jar fr.m2.lsds.App 7 /input/highestRatedMovieNameByUserId.txt /output`
- Save the result and upload it to the `input` folder.
  - `hdfs dfs -put ./hadoop/labs/movieLikeCount.txt /input`
- Make sure to delete the `output` before running the next commands otherwise you will have an error.
- `hadoop jar ./hadoop/labs/Movie-1.0-SNAPSHOT.jar fr.m2.lsds.App 8 /input/movieLikeCount.txt /output`

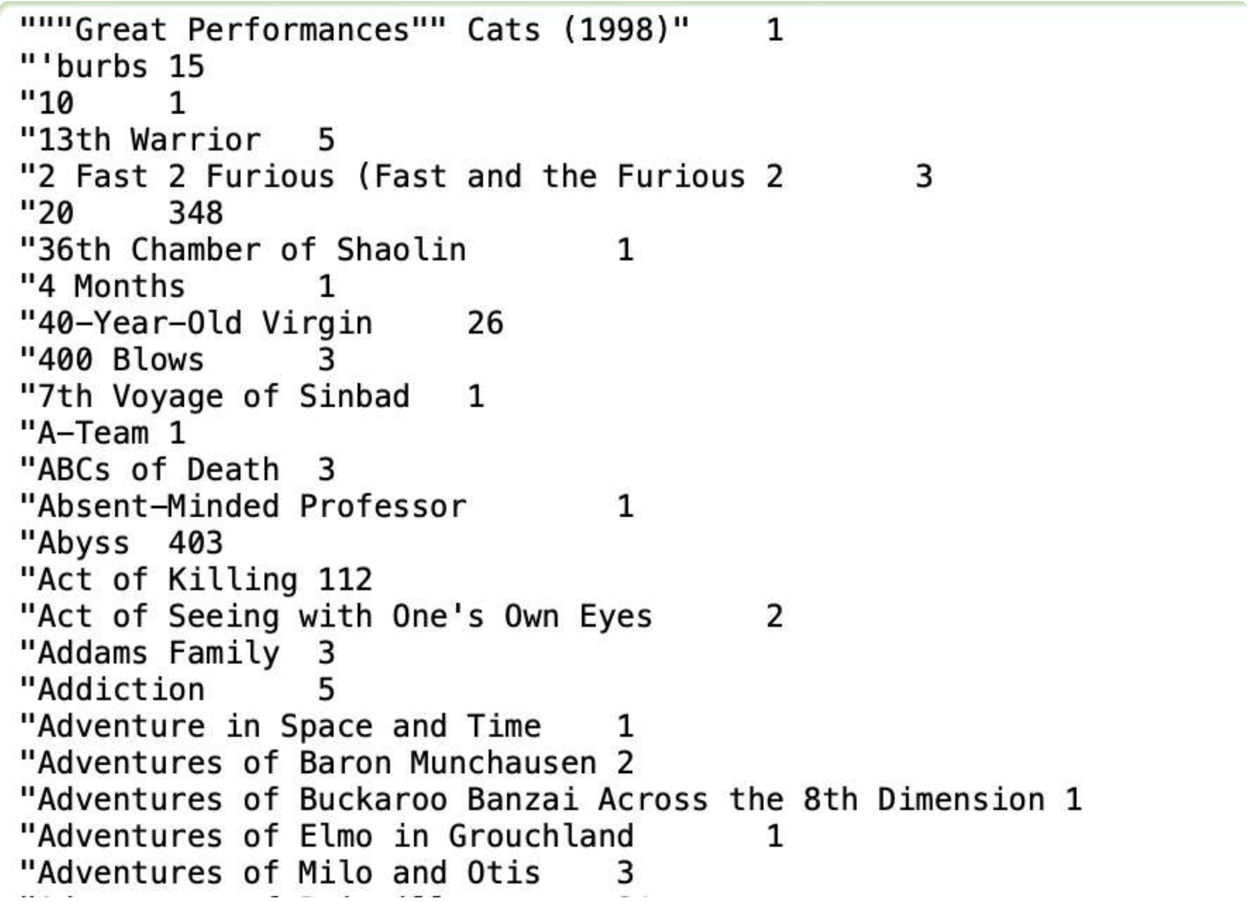
First job

You should see something like this in the terminal :



```
root@d875dff295f9:/# hdfs dfs -put ./hadoop/labs/highestRatedMovieNameByUserId.txt /input
2024-12-21 18:37:33,530 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
root@d875dff295f9:/# clear
root@d875dff295f9:/# hadoop jar ./hadoop/labs/Movie-1.0-SNAPSHOT.jar fr.m2.lsds.App 7 /input/highestRatedMovieNameByUserId.txt /output
2024-12-21 18:39:31,331 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/192.168.97.6:8032
2024-12-21 18:39:31,533 INFO client.AHSProxy: Connecting to Application History server at historyserver/192.168.97.4:10200
2024-12-21 18:39:31,763 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1734804
388373_0004
2024-12-21 18:39:31,860 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-12-21 18:39:31,976 INFO input.FileInputFormat: Total input files to process : 1
2024-12-21 18:39:32,029 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-12-21 18:39:32,498 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-12-21 18:39:32,939 INFO mapreduce.JobSubmitter: number of splits:1
2024-12-21 18:39:33,139 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-12-21 18:39:33,183 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1734804388373_0004
2024-12-21 18:39:33,183 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-12-21 18:39:33,356 INFO conf.Configuration: resource-types.xml not found
2024-12-21 18:39:33,357 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-12-21 18:39:33,666 INFO impl.YarnClientImpl: Submitted application application_1734804388373_0004
2024-12-21 18:39:33,697 INFO mapreduce.Job: The url to track the job: http://resourcemanager:8088/proxy/application_1734804388373_0004/
2024-12-21 18:39:33,698 INFO mapreduce.Job: Running job: job_1734804388373_0004
2024-12-21 18:39:42,037 INFO mapreduce.Job: Job job_1734804388373_0004 running in uber mode : false
2024-12-21 18:39:42,041 INFO mapreduce.Job:  map 0% reduce 0%
2024-12-21 18:39:48,185 INFO mapreduce.Job:  map 100% reduce 0%
2024-12-21 18:39:54,283 INFO mapreduce.Job:  map 100% reduce 100%
2024-12-21 18:39:57,412 INFO mapreduce.Job: Job job_1734804388373_0004 completed successfully
2024-12-21 18:39:57,514 INFO mapreduce.Job: Counters: 54
        File System Counters
```

And have something like this in the `output` folder :

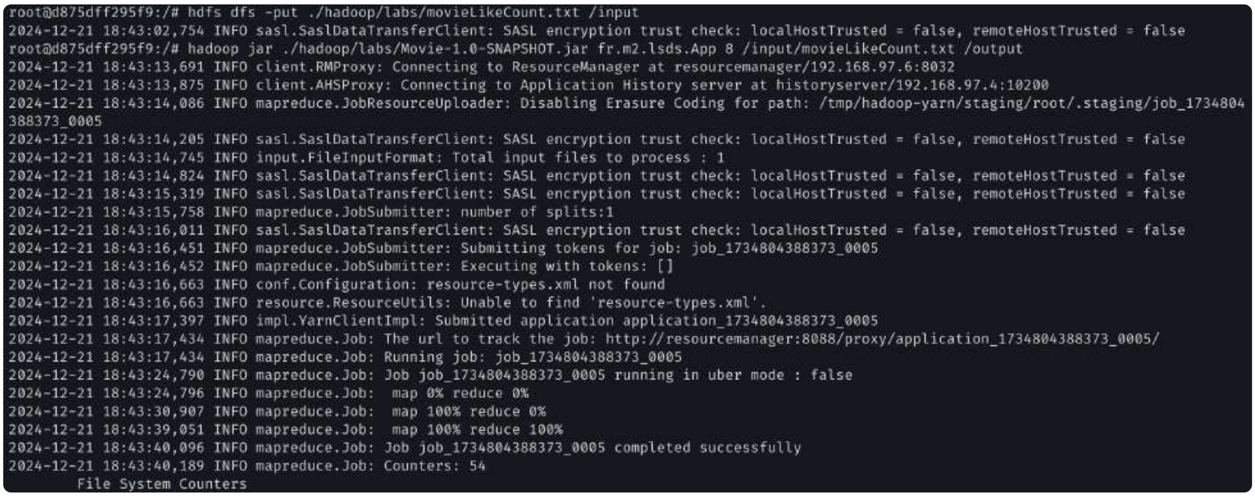


```
"""Great Performances"" Cats (1998)"    1
"'burbs 15
"10     1
"13th Warrior   5
"2 Fast 2 Furious (Fast and the Furious 2       3
"20     348
"36th Chamber of Shaolin        1
"4 Months       1
"40-Year-Old Virgin     26
"400 Blows      3
"7th Voyage of Sinbad   1
"A-Team 1
"ABCs of Death  3
"Absent-Minded Professor        1
"Abyss  403
"Act of Killing 112
"Act of Seeing with One's Own Eyes      2
"Addams Family  3
"Addiction      5
"Adventure in Space and Time    1
"Adventures of Baron Munchausen 2
"Adventures of Buckaroo Banzai Across the 8th Dimension 1
"Adventures of Elmo in Grouchland       1
"Adventures of Milo and Otis    3
```

File results are available in the `Movie/results/movieLikeCount.txt` file.
Code is available in the `Movie/src/main/java/fr/m2/lsds/project/job1` folder.

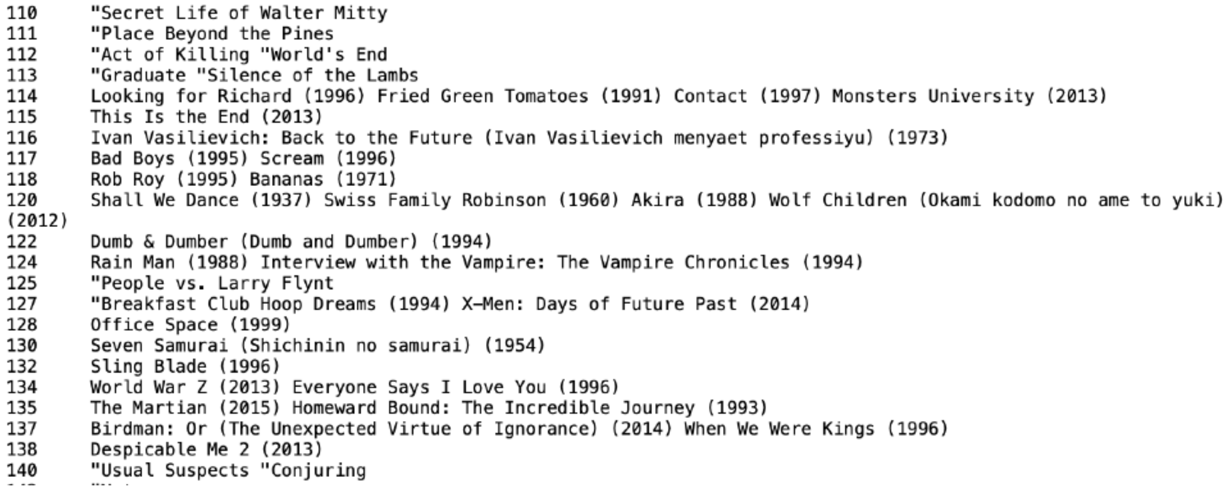
Second job

You should see something like this in the terminal :



```
root@d875dff295f9:/# hdfs dfs -put ./hadoop/labs/movieLikeCount.txt /input
2024-12-21 18:43:02,754 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
root@d875dff295f9:/# hadoop jar ./hadoop/labs/Movie-1.0-SNAPSHOT.jar fr.m2.lsds.App 8 /input/movieLikeCount.txt /output
2024-12-21 18:43:13,691 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/192.168.97.6:8032
2024-12-21 18:43:13,875 INFO client.AHSProxy: Connecting to Application History server at historyserver/192.168.97.4:10200
2024-12-21 18:43:14,086 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1734804
388373_0005
2024-12-21 18:43:14,205 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-12-21 18:43:14,745 INFO input.FileInputFormat: Total input files to process : 1
2024-12-21 18:43:14,824 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-12-21 18:43:15,319 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-12-21 18:43:15,758 INFO mapreduce.JobSubmitter: number of splits:1
2024-12-21 18:43:16,011 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-12-21 18:43:16,451 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1734804388373_0005
2024-12-21 18:43:16,452 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-12-21 18:43:16,663 INFO conf.Configuration: resource-types.xml not found
2024-12-21 18:43:16,663 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-12-21 18:43:17,397 INFO impl.YarnClientImpl: Submitted application application_1734804388373_0005
2024-12-21 18:43:17,434 INFO mapreduce.Job: The url to track the job: http://resourcemanager:8088/proxy/application_1734804388373_0005/
2024-12-21 18:43:17,434 INFO mapreduce.Job: Running job: job_1734804388373_0005
2024-12-21 18:43:24,790 INFO mapreduce.Job: Job job_1734804388373_0005 running in uber mode : false
2024-12-21 18:43:24,796 INFO mapreduce.Job:  map 0% reduce 0%
2024-12-21 18:43:30,907 INFO mapreduce.Job:  map 100% reduce 0%
2024-12-21 18:43:39,051 INFO mapreduce.Job:  map 100% reduce 100%
2024-12-21 18:43:40,096 INFO mapreduce.Job: Job job_1734804388373_0005 completed successfully
2024-12-21 18:43:40,189 INFO mapreduce.Job: Counters: 54
        File System Counters
```

And have something like this in the `output` folder (this is the middle of the file to view multiple movies on the same count and multiple counts):



```
110     "Secret Life of Walter Mitty
111     "Place Beyond the Pines
112     "Act of Killing "World's End
113     "Graduate "Silence of the Lambs
114     Looking for Richard (1996) Fried Green Tomatoes (1991) Contact (1997) Monsters University (2013)
115     This Is the End (2013)
116     Ivan Vasilievich: Back to the Future (Ivan Vasilievich menyaet professiyu) (1973)
117     Bad Boys (1995) Scream (1996)
118     Rob Roy (1995) Bananas (1971)
120     Shall We Dance (1937) Swiss Family Robinson (1960) Akira (1988) Wolf Children (Okami kodomo no ame to yuki)
(2012)
122     Dumb & Dumber (Dumb and Dumber) (1994)
124     Rain Man (1988) Interview with the Vampire: The Vampire Chronicles (1994)
125     "People vs. Larry Flynt
127     "Breakfast Club Hoop Dreams (1994) X-Men: Days of Future Past (2014)
128     Office Space (1999)
130     Seven Samurai (Shichinin no samurai) (1954)
132     Sling Blade (1996)
134     World War Z (2013) Everyone Says I Love You (1996)
135     The Martian (2015) Homeward Bound: The Incredible Journey (1993)
137     Birdman: Or (The Unexpected Virtue of Ignorance) (2014) When We Were Kings (1996)
138     Despicable Me 2 (2013)
140     "Usual Suspects "Conjuring
```

File results are available in the `Movie/results/movieLikeCountGroupByCount.txt` file.
Code is available in the `Movie/src/main/java/fr/m2/lsds/project/job2` folder.