“Hadoop Project” - Report

Github: https://github.com/Kirill2002/HadoopMapReduceProject

Solutions:

Task 1:
Runs two jobs. First job takes ratings.csv as input and outputs a file with the highest rated movieId per each userId. Second job has two mappers. First mapper simply reads movies.csv file and outputs in the following format: (key: movieId, value: “0,movieTitle”), where 0 means that this output was made by the first mapper. The second mapper reads the output of the first job and outputs in the following format: (key: movieId, value: “1,userId”). The reducer runs through the values at most 2 times. First time it looks for the title from the first mapper and then it runs second time to produce outputs for each value from the second mapper.

Task 2:
Runs two jobs (but uses the output of the first task). First job counts the number of likes per movie and outputs it in the following format: (key: movieTitle, value: numberOfLikes). The second job’s mapper reads the output of the first job and outputs: (key: numberOfLikes, value: movieTitle). The reducer then builds the list of movies for each number of likes. It is sorted in ascending order because hadoop’s implementation of map reduce sorts keys before reduce phase (it is needed to make sure that one key is processed by one reducer)

Preparation:

1. Open docker-compose.yml file and replace /home/user/Docker/Labs/Local at 13th line to your shared volume path
2. Open terminal in **Docker** folder
3. Run:

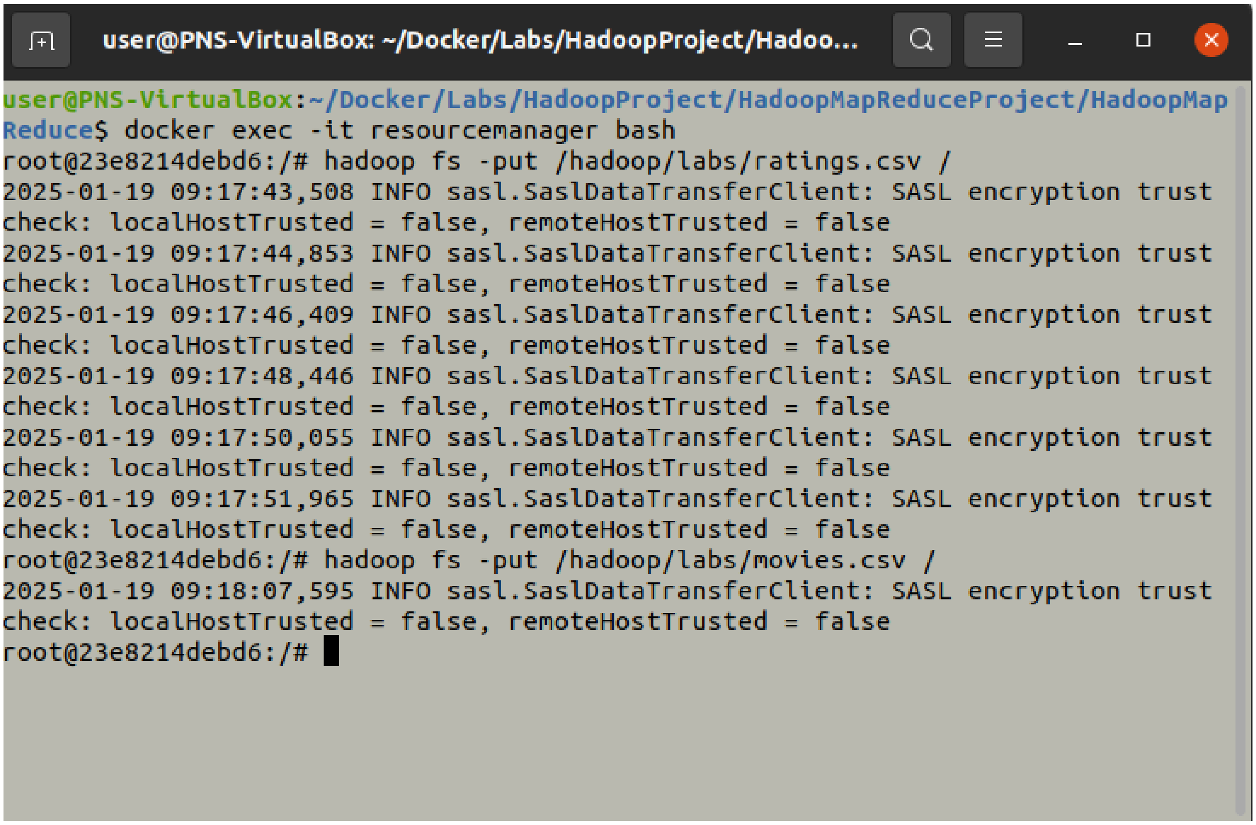
   docker compose up

4. Put input files (ratings.csv and movies.csv) to your shared volume folder
5. Run:

```
docker exec -it resourcemanager bash
```

6. Add files ratings.csv and movies.csv to hadoop distributed filesystem at “/” (see Fig. 1):

```
hadoop fs -put /hadoop/labs/ratings.csv /
hadoop fs -put /hadoop/labs/movies.csv /
```



```
user@PNS-VirtualBox:~/Docker/Labs/HadoopProject/HadoopMapReduceProject/HadoopMap
Reduce$ docker exec -it resourcemanager bash
root@23e8214debd6:/# hadoop fs -put /hadoop/labs/ratings.csv /
2025-01-19 09:17:43,508 INFO sasl.SaslDataTransferClient: SASL encryption trust
check: localHostTrusted = false, remoteHostTrusted = false
2025-01-19 09:17:44,853 INFO sasl.SaslDataTransferClient: SASL encryption trust
check: localHostTrusted = false, remoteHostTrusted = false
2025-01-19 09:17:46,409 INFO sasl.SaslDataTransferClient: SASL encryption trust
check: localHostTrusted = false, remoteHostTrusted = false
2025-01-19 09:17:48,446 INFO sasl.SaslDataTransferClient: SASL encryption trust
check: localHostTrusted = false, remoteHostTrusted = false
2025-01-19 09:17:50,055 INFO sasl.SaslDataTransferClient: SASL encryption trust
check: localHostTrusted = false, remoteHostTrusted = false
2025-01-19 09:17:51,965 INFO sasl.SaslDataTransferClient: SASL encryption trust
check: localHostTrusted = false, remoteHostTrusted = false
root@23e8214debd6:/# hadoop fs -put /hadoop/labs/movies.csv /
2025-01-19 09:18:07,595 INFO sasl.SaslDataTransferClient: SASL encryption trust
check: localHostTrusted = false, remoteHostTrusted = false
root@23e8214debd6:/# 
```

Figure 1 - Adding files to HDFS.

Build project:

1. Open terminal in HadoopMapReduce folder
2. Run:

   ```
   mvn package
   ```

3. Run (replace to your shared volume path):

   ```
   cp target/HadoopMapReduce-1.0-SNAPSHOT.jar /path/to/shared/volume
   ```

Running project:

1. Run:

   ```
   docker exec -it resourcemanager bash
   ```

2. All further commands should be executed in this terminal

Task 1:

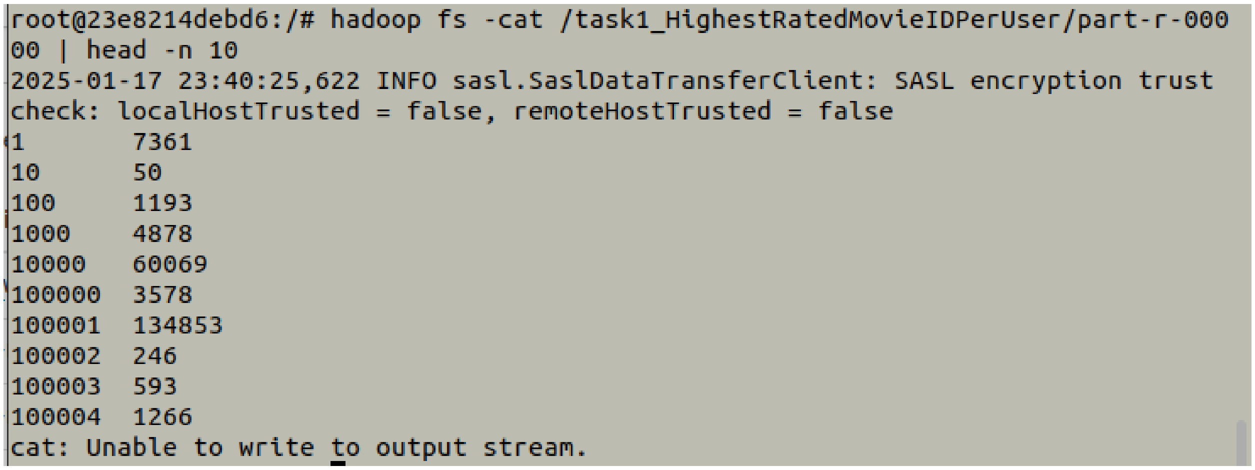
1. Run the command and “cross your fingers”:

   ```
   hadoop jar /hadoop/labs/HadoopMapReduce-1.0-SNAPSHOT.jar
   lsds.project.Task1
   ```

2. The answer to the first part of task1 (“The highest rated movieID per user”) can be found in file in hadoop file system: /task1_HighestRatedMovieIDPerUser/part-r-00000
3. To check 10 first outputs you can run the following command (see Fig. 2):

   ```
   hadoop fs -cat /task1_HighestRatedMovieIDPerUser/part-r-00000 | head -n
   10
   ```

The format is: *userId   movieId*

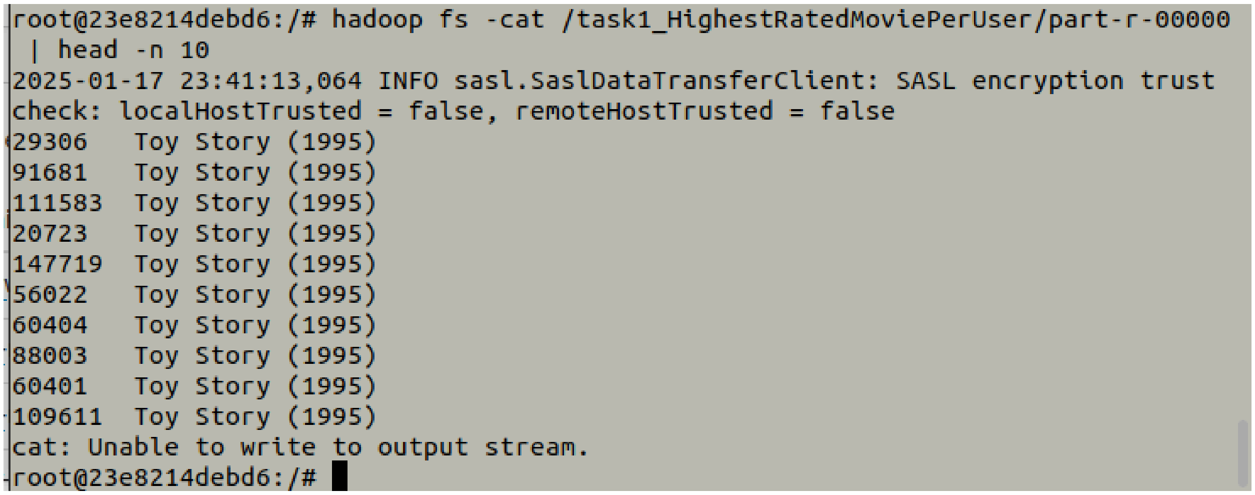


```
root@23e8214debd6:/# hadoop fs -cat /task1_HighestRatedMovieIDPerUser/part-r-000
00 | head -n 10
2025-01-17 23:40:25,622 INFO sasl.SaslDataTransferClient: SASL encryption trust
check: localHostTrusted = false, remoteHostTrusted = false
1       7361
10      50
100     1193
1000    4878
10000   60069
100000  3578
100001  134853
100002  246
100003  593
100004  1266
cat: Unable to write to output stream.
```

Figure 2 - 10 first results of Task 1 part 1.

4. The **final answer to Task1** is at /task1_HighestRatedMoviePerUser/part-r-00000
5. To check 10 first outputs you can run the following command (see Fig. 3):

hadoop fs -cat /task1_HighestRatedMoviePerUser/part-r-00000 | head -n 10



```
root@23e8214debd6:/# hadoop fs -cat /task1_HighestRatedMoviePerUser/part-r-00000
 | head -n 10
2025-01-17 23:41:13,064 INFO sasl.SaslDataTransferClient: SASL encryption trust
check: localHostTrusted = false, remoteHostTrusted = false
29306   Toy Story (1995)
91681   Toy Story (1995)
111583  Toy Story (1995)
20723   Toy Story (1995)
147719  Toy Story (1995)
56022   Toy Story (1995)
60404   Toy Story (1995)
88003   Toy Story (1995)
60401   Toy Story (1995)
109611  Toy Story (1995)
cat: Unable to write to output stream.
root@23e8214debd6:/#
```

Figure 3 - 10 first results of Task 1.

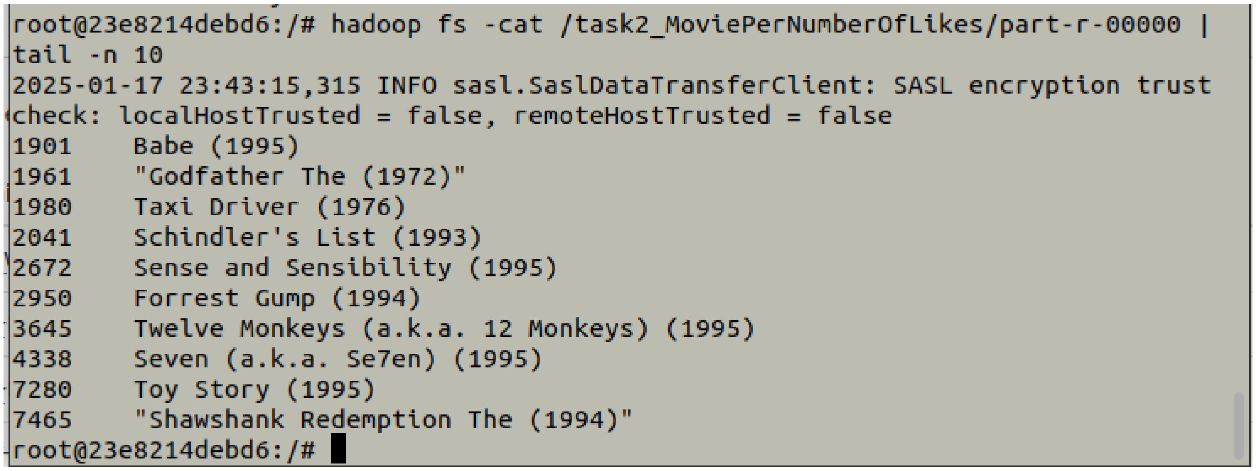
Task 2:

1. **Make sure to run task 1 before**, as task 2 is going to use outputs of task 1
2. Run:

   hadoop jar /hadoop/labs/HadoopMapReduce-1.0-SNAPSHOT.jar lsds.project.Task2

3. The **final answer to Task2** is at /task2_MoviePerNumberOfLikes/part-r-00000
4. To check 10 last outputs (to output 10 lists of most liked movies) run (see Fig.3):

   hadoop fs -cat /task2_MoviePerNumberOfLikes/part-r-00000 | tail -n 10

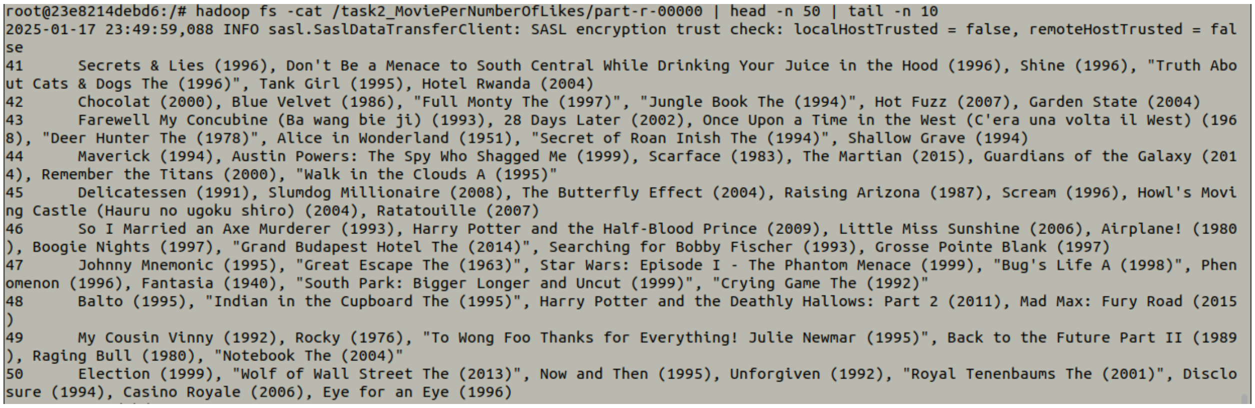


```
root@23e8214debd6:/# hadoop fs -cat /task2_MoviePerNumberOfLikes/part-r-00000 |
tail -n 10
2025-01-17 23:43:15,315 INFO sasl.SaslDataTransferClient: SASL encryption trust
check: localHostTrusted = false, remoteHostTrusted = false
1901    Babe (1995)
1961    "Godfather The (1972)"
1980    Taxi Driver (1976)
2041    Schindler's List (1993)
2672    Sense and Sensibility (1995)
2950    Forrest Gump (1994)
3645    Twelve Monkeys (a.k.a. 12 Monkeys) (1995)
4338    Seven (a.k.a. Se7en) (1995)
7280    Toy Story (1995)
7465    "Shawshank Redemption The (1994)"
root@23e8214debd6:/# 
```

Figure 4 - 10 last results of Task 2.

5. To check outputs 41 - 50 (first outputs have too big lists which makes it hard to read in terminal) so that we can see that movie lists look correct run (see Fig.5) :

   hadoop fs -cat /task2_MoviePerNumberOfLikes/part-r-00000 | head -n 50 | tail -n 10



```
root@23e8214debd6:/# hadoop fs -cat /task2_MoviePerNumberOfLikes/part-r-00000 | head -n 50 | tail -n 10
2025-01-17 23:49:59,088 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = fal
se
41      Secrets & Lies (1996), Don't Be a Menace to South Central While Drinking Your Juice in the Hood (1996), Shine (1996), "Truth Abo
ut Cats & Dogs The (1996)", Tank Girl (1995), Hotel Rwanda (2004)
42      Chocolat (2000), Blue Velvet (1986), "Full Monty The (1997)", "Jungle Book The (1994)", Hot Fuzz (2007), Garden State (2004)
43      Farewell My Concubine (Ba wang bie ji) (1993), 28 Days Later (2002), Once Upon a Time in the West (C'era una volta il West) (196
8), "Deer Hunter The (1978)", Alice in Wonderland (1951), "Secret of Roan Inish The (1994)", Shallow Grave (1994)
44      Maverick (1994), Austin Powers: The Spy Who Shagged Me (1999), Scarface (1983), The Martian (2015), Guardians of the Galaxy (201
4), Remember the Titans (2000), "Walk in the Clouds A (1995)"
45      Delicatessen (1991), Slumdog Millionaire (2008), The Butterfly Effect (2004), Raising Arizona (1987), Scream (1996), Howl's Movi
ng Castle (Hauru no ugoku shiro) (2004), Ratatouille (2007)
46      So I Married an Axe Murderer (1993), Harry Potter and the Half-Blood Prince (2009), Little Miss Sunshine (2006), Airplane! (1980
), Boogie Nights (1997), "Grand Budapest Hotel The (2014)", Searching for Bobby Fischer (1993), Grosse Pointe Blank (1997)
47      Johnny Mnemonic (1995), "Great Escape The (1963)", Star Wars: Episode I - The Phantom Menace (1999), "Bug's Life A (1998)", Phen
omenon (1996), Fantasia (1940), "South Park: Bigger Longer and Uncut (1999)", "Crying Game The (1992)"
48      Balto (1995), "Indian in the Cupboard The (1995)", Harry Potter and the Deathly Hallows: Part 2 (2011), Mad Max: Fury Road (2015
)
49      My Cousin Vinny (1992), Rocky (1976), "To Wong Foo Thanks for Everything! Julie Newmar (1995)", Back to the Future Part II (1989
), Raging Bull (1980), "Notebook The (2004)"
50      Election (1999), "Wolf of Wall Street The (2013)", Now and Then (1995), Unforgiven (1992), "Royal Tenenbaums The (2001)", Disclo
sure (1994), Casino Royale (2006), Eye for an Eye (1996)
```

Figure 5 - lines 41-50 of result of Task 2.