Report for Large Scale Distributed Systems' Project

Gabriele Genovese

18 January 2025

1 Hadoop Project

This project was about creating complex queries on a distributed filesystem. Using the `hadoop` library for `Java`, the main solution was to create a series of Map/Reduce jobs in order to extract how many times a movie was the favourite among the users. I'll present multiple solution to find out the best approch to the problem.

In order to set up the environment to run the project create the `data` folder, move the unzipped data inside it and use the followings commands:

```
docker compose up -d
docker exec -it namenode /bin/bash
hdfs dfs -mkdir /input
hdfs dfs -put /hadoop/labs/movie-data/movies.csv /input
hdfs dfs -put /hadoop/labs/movie-data/ratings.csv /input
```

N.B.: after running the project, if you want to run it again, there is no need to delete the `output` or the `intermediate_output` folders because they are automatically removed by the program.

1.1 First solution: four jobs

The first solution combine 4 jobs to reach the desired result. It fully use the Map/rerduce pattern philosofy.

1.1.1 Design of the solution

This solution is divided in four jobs:

- the first one is in charge of joining the two file using movieId as a key
- the second one format the data to use userId as a key
- the third one choose a random favorite movie per user
- the fourth compute the frequency and format the output

In the code, there are comments over each function that specify how the input and the output of each map reduce is managed.

1.1.2 Execution

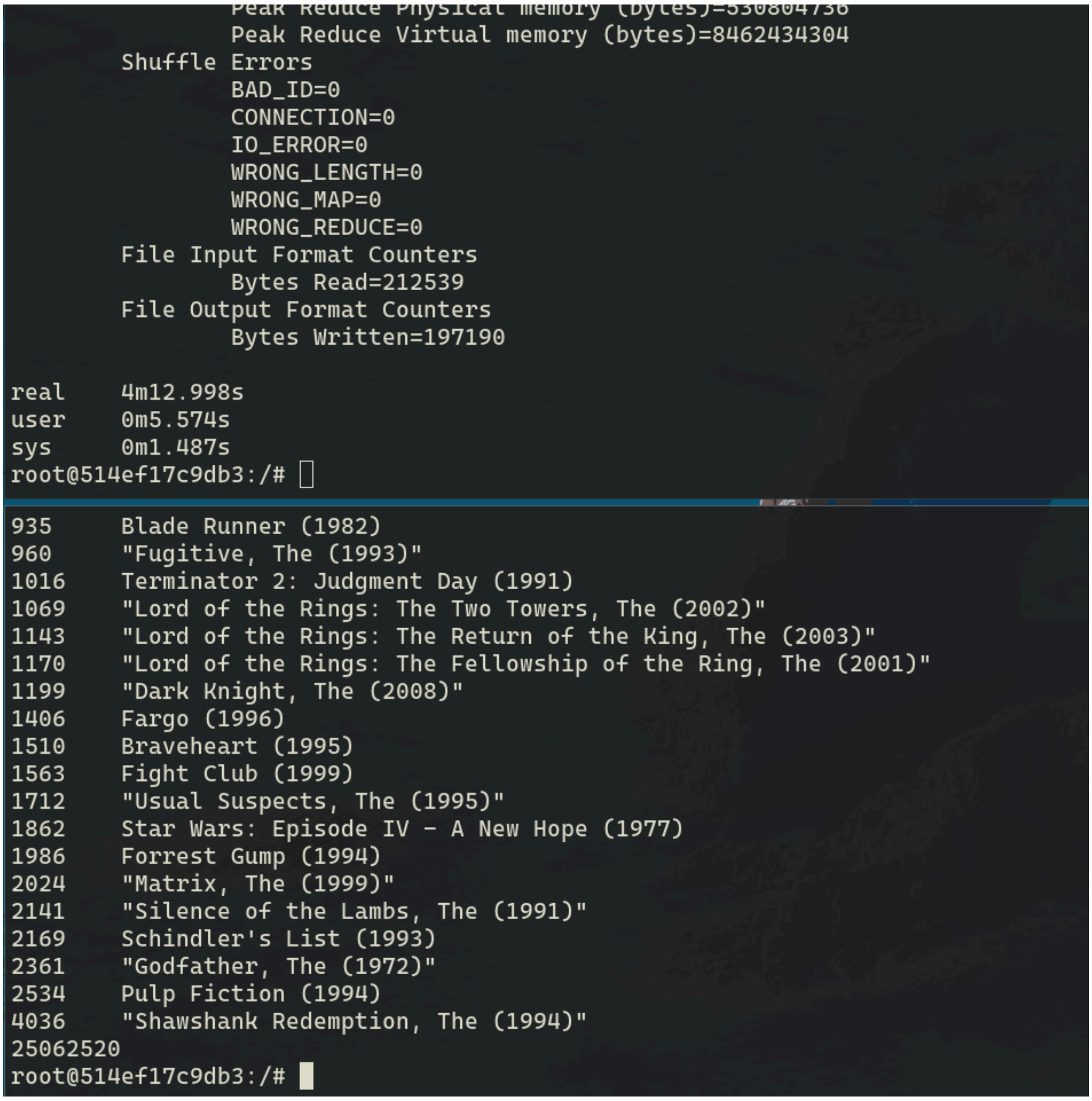
To execute the code, open a new terminal and use the followings commands in the host machine:

```
mvn package
cp target/hadoop-1.0.jar data
```

Then, run this command in the `namenode` container:

```
time hadoop jar /hadoop/labs/hadoop-1.0.jar app.ChainFirst /input /output
```

The program should give the following output:



```
                Peak Reduce Physical memory (bytes)=530804736
                Peak Reduce Virtual memory (bytes)=8462434304
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=212539
        File Output Format Counters
                Bytes Written=197190

real    4m12.998s
user    0m5.574s
sys     0m1.487s
root@514ef17c9db3:/# 
```

```
935     Blade Runner (1982)
960     "Fugitive, The (1993)"
1016    Terminator 2: Judgment Day (1991)
1069    "Lord of the Rings: The Two Towers, The (2002)"
1143    "Lord of the Rings: The Return of the King, The (2003)"
1170    "Lord of the Rings: The Fellowship of the Ring, The (2001)"
1199    "Dark Knight, The (2008)"
1406    Fargo (1996)
1510    Braveheart (1995)
1563    Fight Club (1999)
1712    "Usual Suspects, The (1995)"
1862    Star Wars: Episode IV - A New Hope (1977)
1986    Forrest Gump (1994)
2024    "Matrix, The (1999)"
2141    "Silence of the Lambs, The (1991)"
2169    Schindler's List (1993)
2361    "Godfather, The (1972)"
2534    Pulp Fiction (1994)
4036    "Shawshank Redemption, The (1994)"
25062520
root@514ef17c9db3:/# 
```

Abb. 1: First's solution output showed using `hdfs dfs -cat /output/part-r-00000`

The solution run in about 4 minutes and 13 seconds. The most favoured movie was “The Shawshank Redemption” (1994) with 4036 preference.

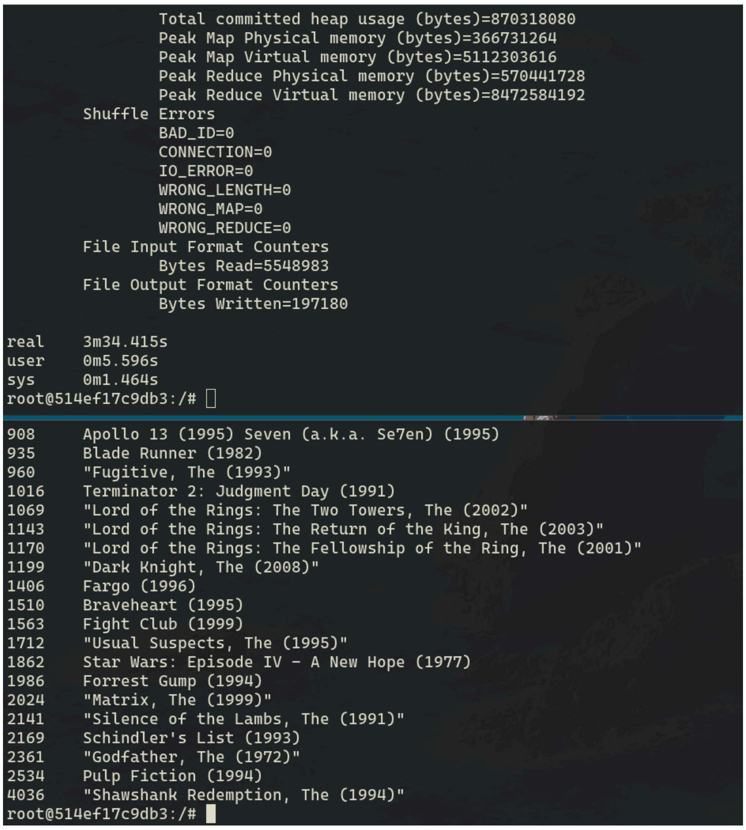
1.2 Other solutions: three and two jobs

It's also possible to execute different kind of solutions with lower jobs. This was done because I noted that creating an entire new job is an heavy operation, so I wanted to explore how the execution time can improve using less jobs.

The two other solutions can be executed using:

```
time hadoop jar /hadoop/labs/hadoop-1.0.jar app.ChainSec /input /output
time hadoop jar /hadoop/labs/hadoop-1.0.jar app.ChainTer /input /output
```

The second solution with three jobs ends in about 3 minutes and 40 minutes. The third solution with two jobs ends in about 2 minutes and 8 minutes.

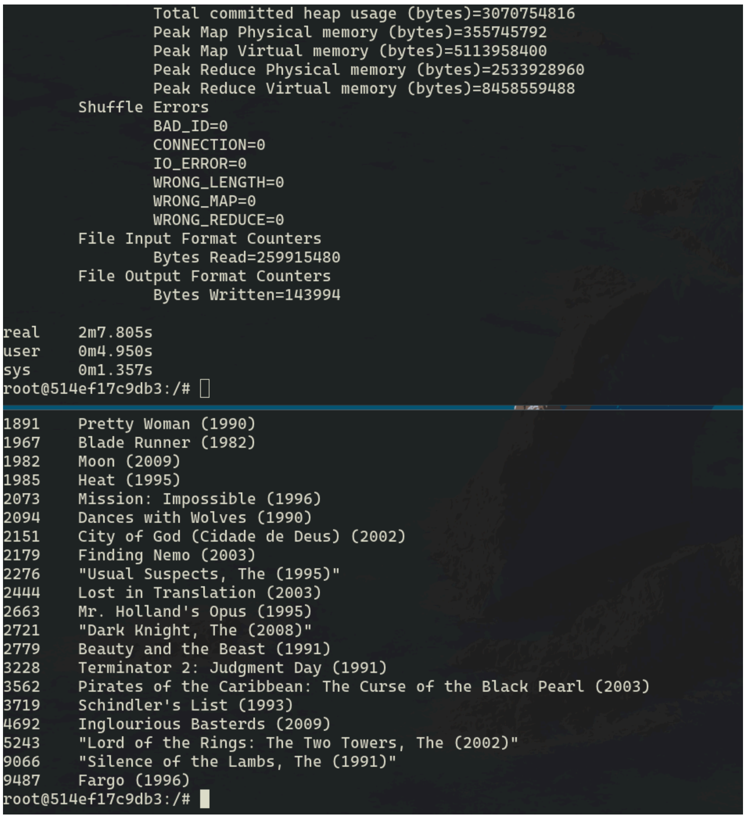


```
                Total committed heap usage (bytes)=870318080
                Peak Map Physical memory (bytes)=366731264
                Peak Map Virtual memory (bytes)=5112303616
                Peak Reduce Physical memory (bytes)=570441728
                Peak Reduce Virtual memory (bytes)=8472584192
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=5548983
        File Output Format Counters
                Bytes Written=197180

real    3m34.415s
user    0m5.596s
sys     0m1.464s
root@514ef17c9db3:/# 
```

```
908     Apollo 13 (1995) Seven (a.k.a. Se7en) (1995)
935     Blade Runner (1982)
960     "Fugitive, The (1993)"
1016    Terminator 2: Judgment Day (1991)
1069    "Lord of the Rings: The Two Towers, The (2002)"
1143    "Lord of the Rings: The Return of the King, The (2003)"
1170    "Lord of the Rings: The Fellowship of the Ring, The (2001)"
1199    "Dark Knight, The (2008)"
1406    Fargo (1996)
1510    Braveheart (1995)
1563    Fight Club (1999)
1712    "Usual Suspects, The (1995)"
1862    Star Wars: Episode IV - A New Hope (1977)
1986    Forrest Gump (1994)
2024    "Matrix, The (1999)"
2141    "Silence of the Lambs, The (1991)"
2169    Schindler's List (1993)
2361    "Godfather, The (1972)"
2534    Pulp Fiction (1994)
4036    "Shawshank Redemption, The (1994)"
root@514ef17c9db3:/# 
```

Abb. 2: Second's solution output



```
                Total committed heap usage (bytes)=3070754816
                Peak Map Physical memory (bytes)=355745792
                Peak Map Virtual memory (bytes)=5113958400
                Peak Reduce Physical memory (bytes)=2533928960
                Peak Reduce Virtual memory (bytes)=8458559488
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=259915480
        File Output Format Counters
                Bytes Written=143994

real    2m7.805s
user    0m4.950s
sys     0m1.357s
root@514ef17c9db3:/# 
```

```
1891    Pretty Woman (1990)
1967    Blade Runner (1982)
1982    Moon (2009)
1985    Heat (1995)
2073    Mission: Impossible (1996)
2094    Dances with Wolves (1990)
2151    City of God (Cidade de Deus) (2002)
2179    Finding Nemo (2003)
2276    "Usual Suspects, The (1995)"
2444    Lost in Translation (2003)
2663    Mr. Holland's Opus (1995)
2721    "Dark Knight, The (2008)"
2779    Beauty and the Beast (1991)
3228    Terminator 2: Judgment Day (1991)
3562    Pirates of the Caribbean: The Curse of the Black Pearl (2003)
3719    Schindler's List (1993)
4692    Inglourious Basterds (2009)
5243    "Lord of the Rings: The Two Towers, The (2002)"
9066    "Silence of the Lambs, The (1991)"
9487    Fargo (1996)
root@514ef17c9db3:/# 
```

Abb. 3: Third's solution output (note that is different from the other two solutions, this is due to the randomness of chosing a user's favorit movie)