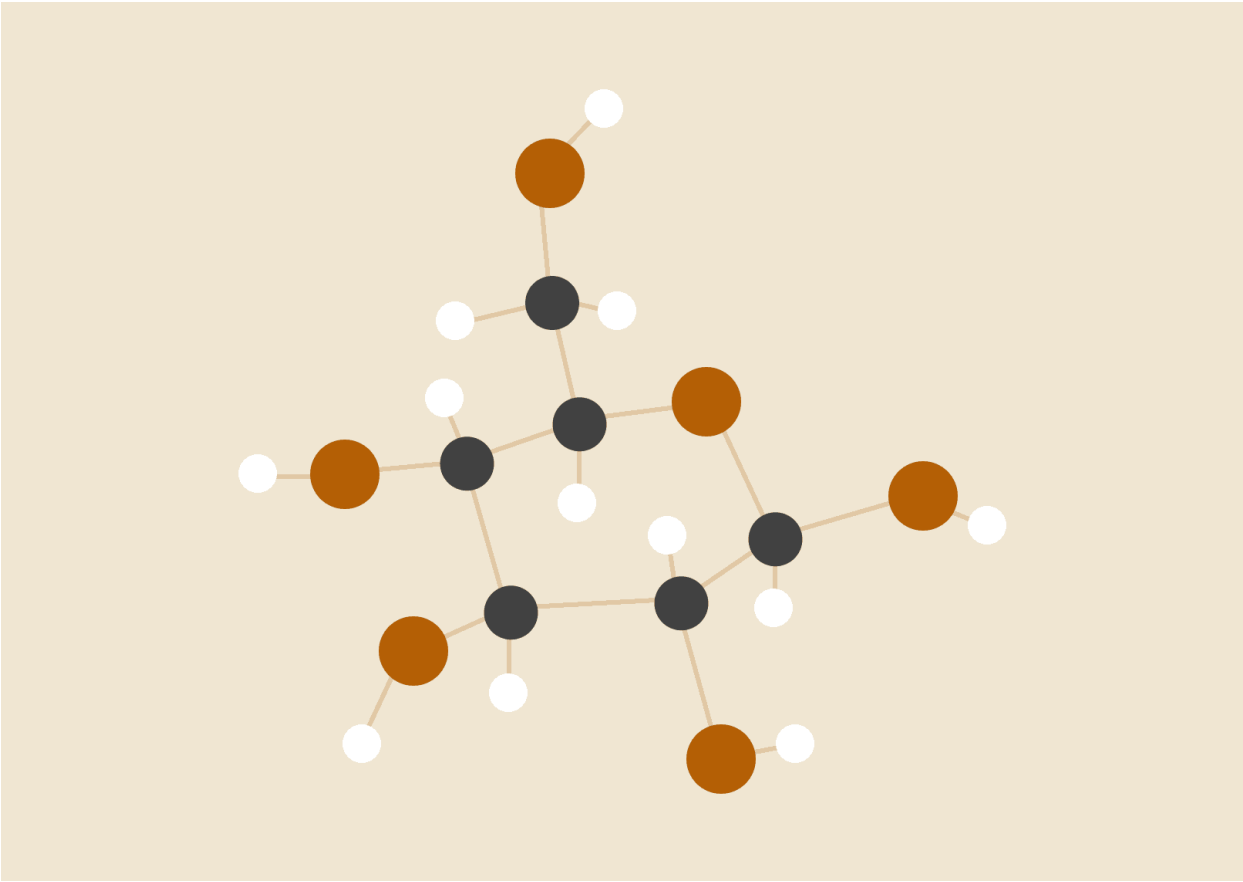
Movie MapReduce

Un projet de traitement de films avec Hadoop MapReduce



Adjo Emmanuelle ADOTE

15.01.2025

SI5 - SSE

INTRODUCTION

Ce document s'inscrit dans le cadre du mini-projet portant sur Hadoop MapReduce. Il a pour but de présenter les étapes nécessaires à l'exécution du projet, ainsi qu'une série de captures montrant l'exécution et les résultats du code produit.

HYPOTHESES

Le projet a été préalablement cloné depuis l'URL : https://github.com/EmmanuelleAD/MovieMapRed, et vous êtes à la racine du projet.

Vous disposez du dossier *ml-25ml* et des fichiers qu'il contient.

PROCEDURE D'EXECUTION

1. Lancez la commande pour générer le jar ***mvn clean package***

2. Lancez les conteneurs avec ***docker compose up -d***
3. Ouvrez un terminal dans le conteneur namenode avec ***docker exec -it namenode bash***
4. Créez un répertoire input sur hdfs pour vos fichiers à passer en paramètre ***hdfs dfs -mkdir -p /input***
5. Ouvrez un autre terminal et copiez les fichiers movies.csv et ratings.csv de *ml-25ml* sur le conteneur namenode :
   docker cp "<path>/movies.csv" namenode:/tmp/movies.csv

   docker cp "<path>/ratings.csv" namenode:/tmp/ratings.csv

6. Retournez dans le terminal ouvert dans namenode et copiez ces fichiers sur hdfs avec

   hdfs dfs -put /tmp/ratings.csv /input/ratings.csv
   hdfs dfs -put /tmp/movies.csv /input/movies.csv

7. Exécutez le premier job MapReduce avec :
   hadoop jar /hadoop/labs/target/MovieMapRed-1.0-SNAPSHOT.jar org.example.HighestRatedMovieNamePerUserId /input/ratings.csv /input/movies.csv <result1>

8. Analysez les résultats intermédiaires (Optionnel)
   hdfs dfs -cat <result1>/part-*

9. Lancer le second job avec la sortie du précédent job

   hadoop jar /hadoop/labs/target/MovieMapRed-1.0-SNAPSHOT.jar org.example.HighestRatedMovieCount <result1> <outputPath>

10. Affichez les résultats avec
    hdfs dfs -cat <outputPath>/part-*

RESULTATS:

1. Exécution de l'étape 1 et 2

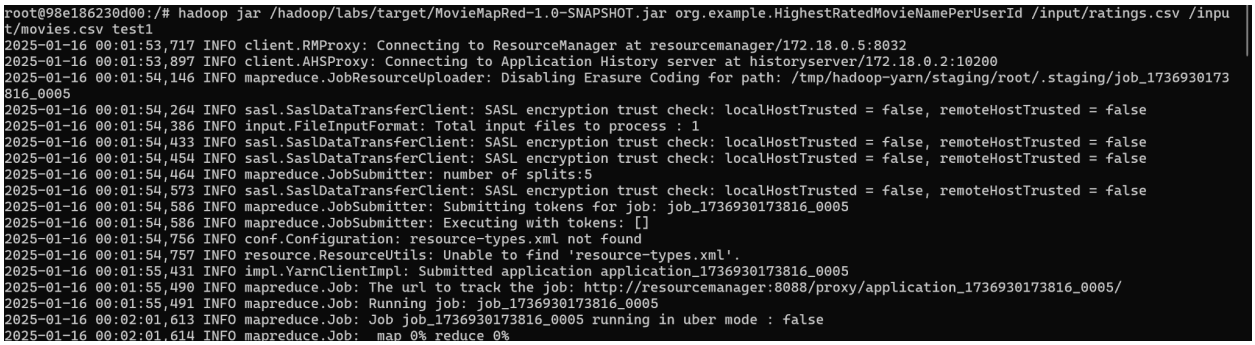
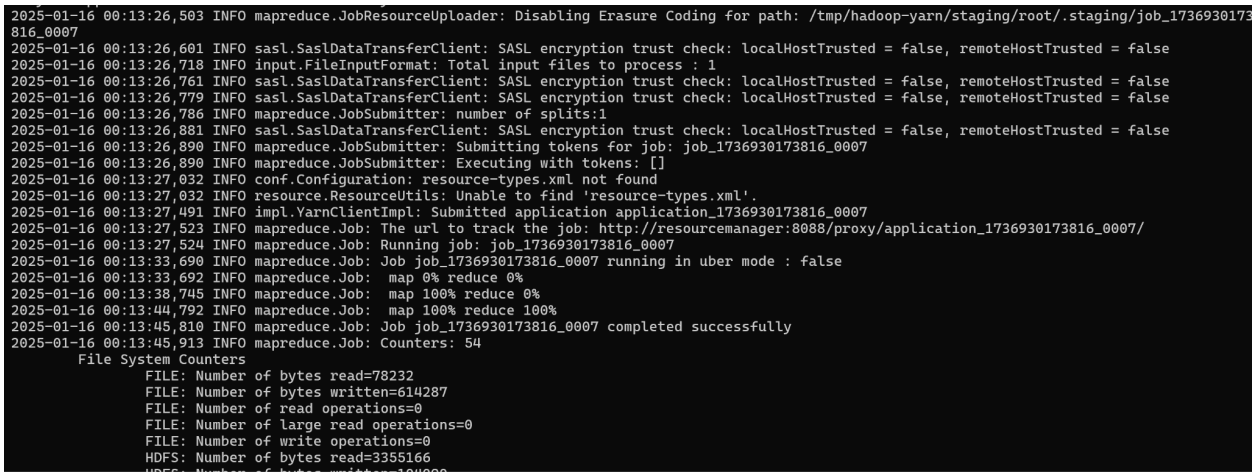


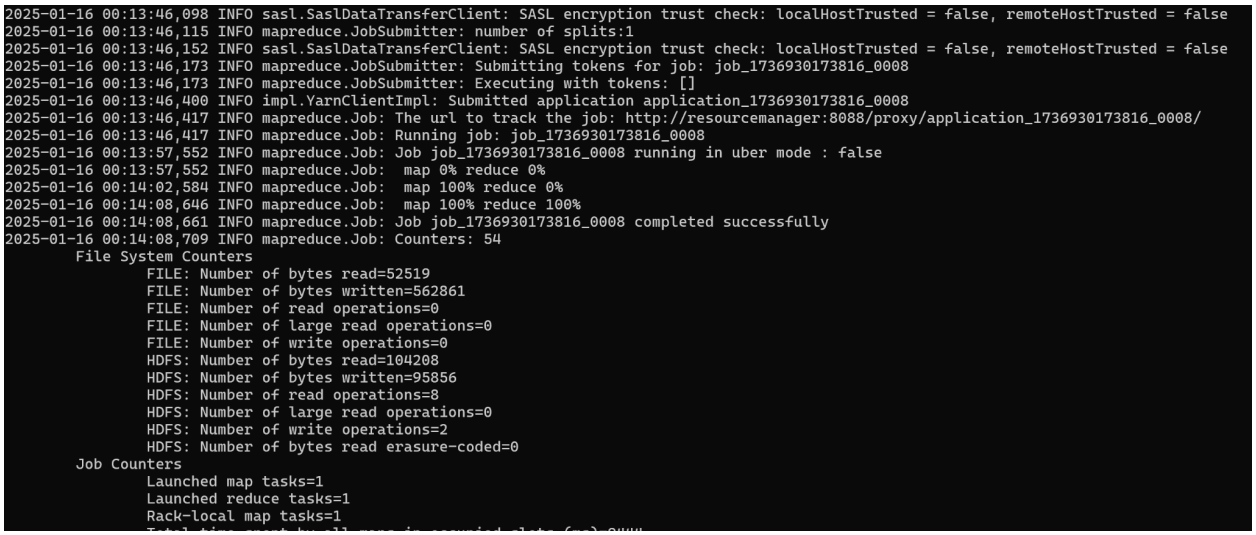
Figure montrant l'exécution du premier job pour obtenir les noms des films les mieux notés par utilisateur

```
2025-01-16 00:03:22,535 INFO mapreduce.Job: Job job_1736930173816_0005 completed successfully
2025-01-16 00:03:22,684 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=187008720
                FILE: Number of bytes written=281848362
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=678277886
                HDFS: Number of bytes written=1681517
                HDFS: Number of read operations=20
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=5
                Launched reduce tasks=1
                Rack-local map tasks=5
                Total time spent by all maps in occupied slots (ms)=366912
                Total time spent by all reduces in occupied slots (ms)=328192
                Total time spent by all map tasks (ms)=91728
                Total time spent by all reduce tasks (ms)=41024
                Total vcore-milliseconds taken by all map tasks=91728
                Total vcore-milliseconds taken by all reduce tasks=41024
                Total megabyte-milliseconds taken by all map tasks=375717888
                Total megabyte-milliseconds taken by all reduce tasks=336068608
        Map-Reduce Framework
                Map input records=25000096
                Map output records=25000095
                Map output bytes=384149336
                Map output materialized bytes=93464869
                Input split bytes=515
                Combine input records=0
                Combine output records=0
                Reduce input groups=162541
                Reduce shuffle bytes=93464869
                Reduce input records=25000095
```

```
2025-01-16 00:03:22,928 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2025-01-16 00:03:22,940 INFO mapreduce.JobSubmitter: number of splits:2
2025-01-16 00:03:22,966 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2025-01-16 00:03:22,984 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1736930173816_0006
2025-01-16 00:03:22,985 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-01-16 00:03:23,219 INFO impl.YarnClientImpl: Submitted application application_1736930173816_0006
2025-01-16 00:03:23,227 INFO mapreduce.Job: The url to track the job: http://resourcemanager:8088/proxy/application_1736930173816_0006/
2025-01-16 00:03:23,227 INFO mapreduce.Job: Running job: job_1736930173816_0006
2025-01-16 00:03:33,337 INFO mapreduce.Job: Job job_1736930173816_0006 running in uber mode : false
2025-01-16 00:03:33,338 INFO mapreduce.Job:  map 0% reduce 0%
2025-01-16 00:03:39,444 INFO mapreduce.Job:  map 100% reduce 0%
2025-01-16 00:03:45,483 INFO mapreduce.Job:  map 100% reduce 100%
2025-01-16 00:03:45,492 INFO mapreduce.Job: Job job_1736930173816_0006 completed successfully
2025-01-16 00:03:45,519 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=1715865
                FILE: Number of bytes written=4042591
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=4720185
                HDFS: Number of bytes written=3355052
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
```

Figures montrant l'exécution réussie des 2 jobs pour obtenir les noms des films les mieux notés par utilisateur





Figures montrant l’exécution réussie des 2 jobs pour obtenir le groupement des films par popularité auprès des utilisateurs

2. Noms des films les mieux notés par utilisateur

```
105389  "Misérables
102840  "Misérables
97500   "Misérables
58832   Big Night (1996)
105444  Big Night (1996)
68151   Big Night (1996)
53834   Big Night (1996)
16765   Big Night (1996)
124265  Big Night (1996)
102781  Big Night (1996)
7942    Big Night (1996)
15298   Big Night (1996)
1109    Big Night (1996)
130735  Big Night (1996)
137225  Big Night (1996)
154438  Big Night (1996)
14175   Big Night (1996)
104891  Big Night (1996)
16736   Big Night (1996)
158625  Big Night (1996)
33084   Big Night (1996)
100317  Big Night (1996)
73051   Big Night (1996)
142315  Big Night (1996)
70388   Big Night (1996)
70684   John Dies at the End (2012)
81020   Last Man Standing (1996)
7427    Last Man Standing (1996)
64695   Last Man Standing (1996)
20219   Last Man Standing (1996)
8802    It's Such a Beautiful Day (2012)
10817   It's Such a Beautiful Day (2012)
12037   It's Such a Beautiful Day (2012)
104120  Set It Off (1996)
79064   2 Days in the Valley (1996)
93900   2 Days in the Valley (1996)
101288  "Last Stand
88461   Upstream Color (2013)
2653    Upstream Color (2013)
141201  Upstream Color (2013)
```

Figure montrant une partie des films les mieux notés par utilisateur provenant de l'exécution du premier lot de job

3. Groupement des films par la popularité en nombre d'utilisateurs

```
425     Grumpier Old Men (1995)
454     Happy Gilmore (1996)
460     One Flew Over the Cuckoo's Nest (1975)
462     "Fugitive
468     "Dark Knight
479     Monty Python and the Holy Grail (1975)
485     Casablanca (1942)
494     American Beauty (1999)
500     Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981)
561     Sabrina (1995)
588     "Princess Bride
609     "Postman
705     "Lion King
726     Clueless (1995)
728     Terminator 2: Judgment Day (1991)
733     "Lord of the Rings: The Fellowship of the Ring
740     Clerks (1994)
753     Star Wars: Episode V - The Empire Strikes Back (1980)
754     Jumanji (1995)
801     Mr. Holland's Opus (1995)
836     Fargo (1996)
932     Fight Club (1999)
962     "City of Lost Children
986     GoldenEye (1995)
1034    Get Shorty (1995)
1064    "American President
1196    Dead Man Walking (1995)
1215    Blade Runner (1982)
1252    Casino (1995)
1277    "Matrix
1406    "Silence of the Lambs
1436    Leaving Las Vegas (1995)
1901    Babe (1995)
1961    "Godfather
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)
7465    "Shawshank Redemption
root@98e186230d00:/#
```

Figure montrant les derniers films provenant du groupement des films par popularité auprès des utilisateurs

REFERENCES

1. https://github.com/EmmanuelleAD/MovieMapRed
2. ml-25m

3. https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html#Example:_WordCount_v2.0
4. https://copyprogramming.com/howto/hadoop-multiple-inputs