Compte Rendu TP1

GL4

Hadoop et Map Reduce

Binômes

- Oumaima KBOUBI
- Rami Kammoun

Installation

Aprés le téléchargement de l'image docker uploadée sur dockerhub: **liliasfaxi/spark-hadoop:hv-2.7.2** et la création d'un réseau permettant de relier 3 contenaires, on crée ces 3 contenaires et on les lance

```
Command Prompt

C:\Users\ouma>docker run -itd --net=hadoop -p 50070:50070 -p 8088:8088 -p 7077:707
7 -p16010:16010 \ --name hadoop-master --hostname hadoop-master \ liliasfaxi/spark
-hadoop:hv-2.7.2
docker: invalid reference format.
See 'docker run --help'.

C:\Users\ouma>docker run -itd --net=hadoop -p 50070:50070 -p 8088:8088 -p 7077:707
7 -p16010:16010  --name hadoop-master --hostname hadoop-master  liliasfaxi/spark-h
adoop:hv-2.7.2
e48dd7fd041fba2ef106f3d5422bc7cd4a04407cb23df9e2e86dc389757424f6

C:\Users\ouma>docker run -itd --net=hadoop -p 8040:8042  --name hadoop-slave1 --ho
stname hadoop-slave1  liliasfaxi/spark-hadoop:hv-2.7.2
e8c5c0e2567b4ce4778015ff0f529d5201a167091e972d8cbac6aa1ea9e08168

C:\Users\ouma>docker run -itd --net=hadoop -p 8041:8042  --name hadoop-slave2 --ho
stname hadoop-slave2  liliasfaxi/spark-hadoop:hv-2.7.2
080bc05fa32216e8a588fc0075189fd858d131daed2eedb678d43ca5a3259b6b

C:\Users\ouma>_
```

On entre dans le contenaire master pour pouvoir l'utiliser:

```
C:\Users\ouma>docker run -itd --net=hadoop -p 8041:8042  --name hadoop-slave2 --hos
080bc05fa32216e8a588fc0075189fd858d131daed2eedb678d43ca5a3259b6b

C:\Users\ouma>docker exec -it hadoop-master bash
root@hadoop-master:~# ^C
```

Pour commencer les manipulations, il faut lancer Hadoop et yarn

```
root@hadoop-master:~# ./start-hadoop.sh


Starting namenodes on [hadoop-master]
hadoop-master: Warning: Permanently added 'hadoop-master,172.18.0.2' (ECDSA) to the list of known hosts.
hadoop-master: starting namenode, logging to /usr/local/hadoop/logs/hadoop-root-namenode-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-slave2,172.18.0.4' (ECDSA) to the list of known hosts.
hadoop-slave1: starting datanode, logging to /usr/local/hadoop/logs/hadoop-root-datanode-hadoop-slave1.out
hadoop-slave2: starting datanode, logging to /usr/local/hadoop/logs/hadoop-root-datanode-hadoop-slave2.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: Warning: Permanently added '0.0.0.0' (ECDSA) to the list of known hosts.
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-slave2,172.18.0.4' (ECDSA) to the list of known hosts.
hadoop-slave1: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-root-nodemanager-hadoop-slave1.out
hadoop-slave2: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-root-nodemanager-hadoop-slave2.out


root@hadoop-master:~#
```

- Lancer Hadoop-master
- Lancer le namenodes secondaire
- Lancer yarn

Premiers pas avec Hadoop

1. Création d'un répertoire "**input**" dans HDFS
2. Afficher les répertoires/fichier dans root
3. Charger le fichier "**purchases.txt**" dans le répertoire **input**
4. Afficher le contenu du répertoire **input** dans HDFS pour vérifier le chargement du fichier
5. Afficher les dernières lignes du fichier **purchases.txt**

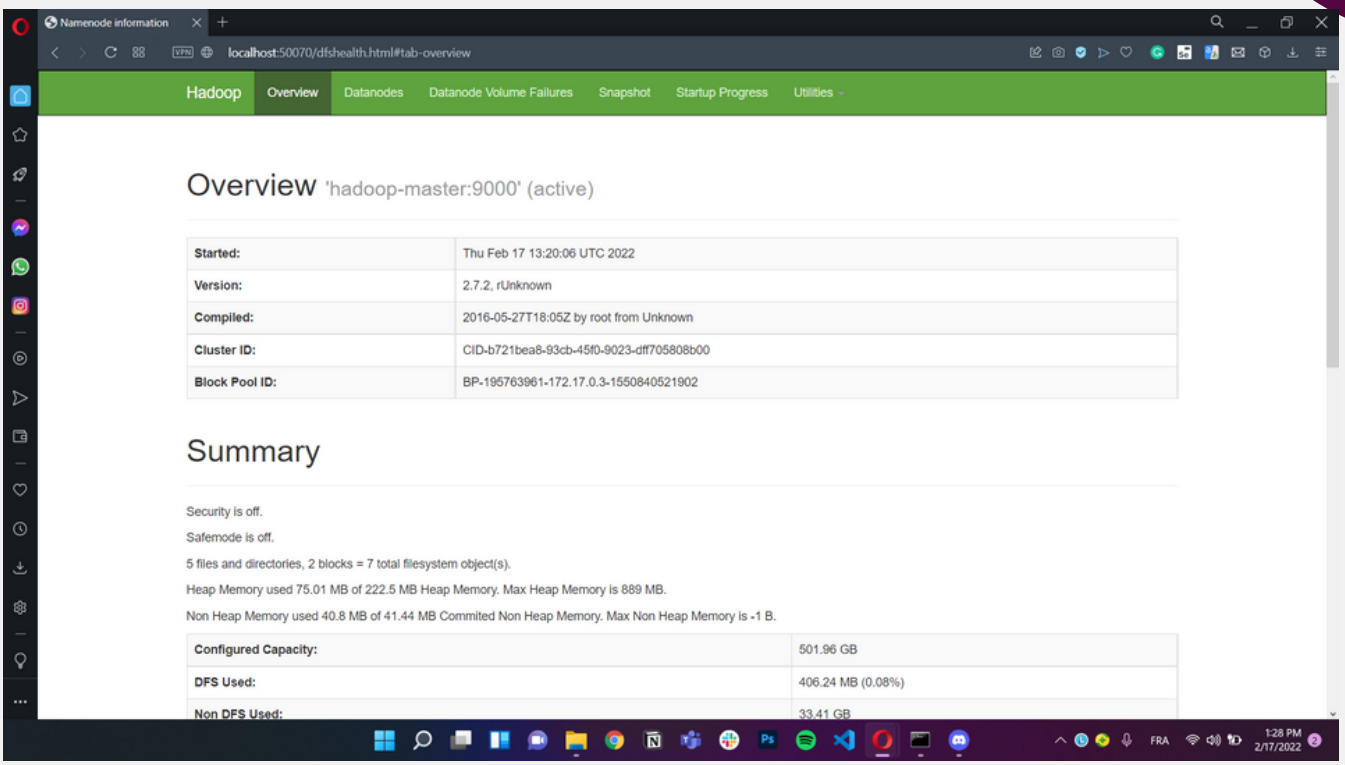
Le résultat de l'exécution de ses commandes:

```
root@hadoop-master:~# hadoop fs -mkdir -p input        1
root@hadoop-master:~# ls        2
hdfs  purchases.txt  purchases2.txt  run-wordcount.sh  start-hadoop.sh  start-kafka-zookeeper.sh
root@hadoop-master:~# hadoop fs -put purchases.txt input        3
root@hadoop-master:~# hadoop fs -ls input        4
Found 1 items
-rw-r--r--   2 root supergroup  211312924 2022-02-17 13:23 input/purchases.txt
root@hadoop-master:~# hadoop fs -tail input/purchases.txt        5
31      17:59   Norfolk Toys    164.34  MasterCard
2012-12-31      17:59   Chula Vista     Music   380.67  Visa
2012-12-31      17:59   Hialeah Toys    115.21  MasterCard
2012-12-31      17:59   Indianapolis    Men's Clothing  158.28  MasterCard
2012-12-31      17:59   Norfolk Garden  414.09  MasterCard
2012-12-31      17:59   Baltimore       DVDs    467.3   Visa
2012-12-31      17:59   Santa Ana       Video Games     144.73  Visa
2012-12-31      17:59   Gilbert Consumer Electronics    354.66  Discover
2012-12-31      17:59   Memphis Sporting Goods  124.79  Amex
2012-12-31      17:59   Chicago Men's Clothing  386.54  MasterCard
2012-12-31      17:59   Birmingham      CDs     118.04  Cash
2012-12-31      17:59   Las Vegas       Health and Beauty       420.46  Amex
2012-12-31      17:59   Wichita Toys    383.9   Cash
2012-12-31      17:59   Tucson  Pet Supplies    268.39  MasterCard
2012-12-31      17:59   Glendale        Women's Clothing        68.05   Amex
2012-12-31      17:59   Albuquerque     Toys    345.7   MasterCard
2012-12-31      17:59   Rochester       DVDs    399.57  Amex
2012-12-31      17:59   Greensboro      Baby    277.27  Discover
2012-12-31      17:59   Arlington       Women's Clothing        134.95  MasterCard
2012-12-31      17:59   Corpus Christi  DVDs    441.61  Discover
root@hadoop-master:~#
```

Interfaces web pour Hadoop

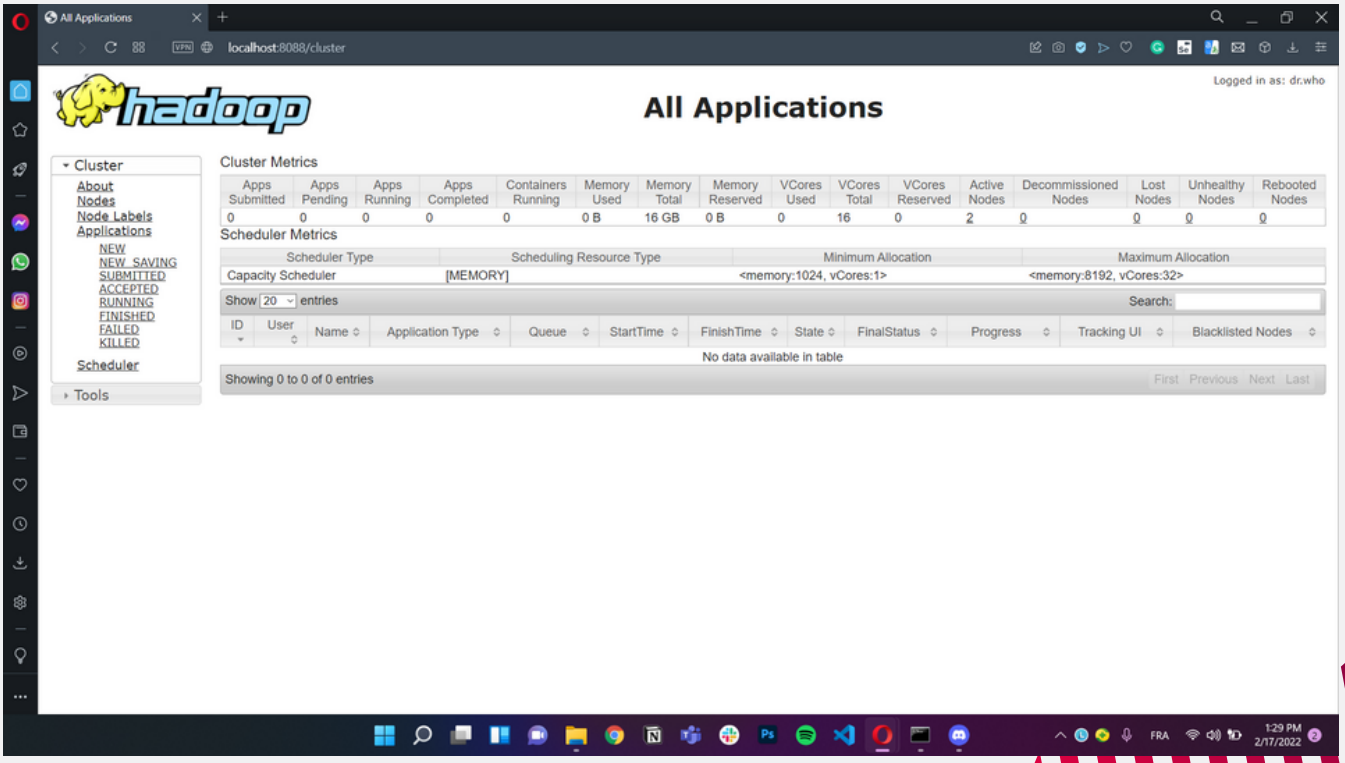
Pour observer le comportement de ses différentes composantes de Hadoop:

- **http://localhost:50070** : affiche les informations de notre namenode



Pour visualiser l'avancement et les résultats de nos Jobs:

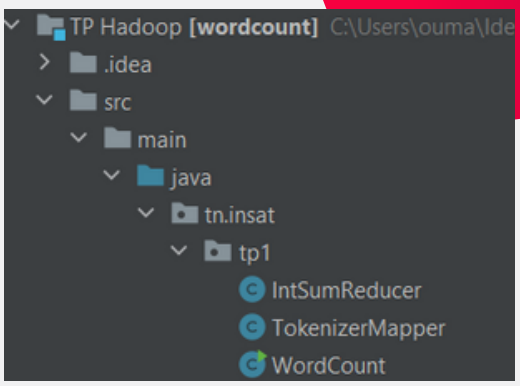
- **http://localhost:8088** : affiche les informations du resource manager de Yarn et visualiser le comportement des différents jobs.

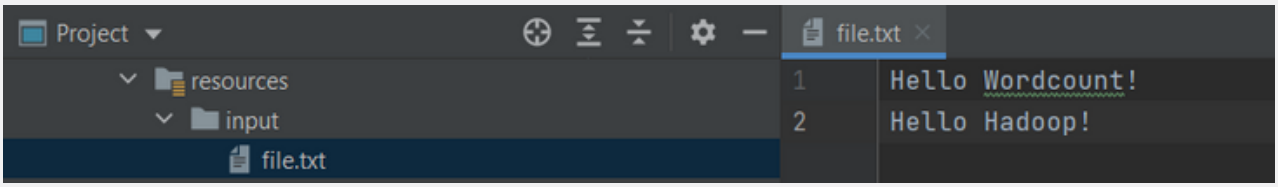


Map Reduce

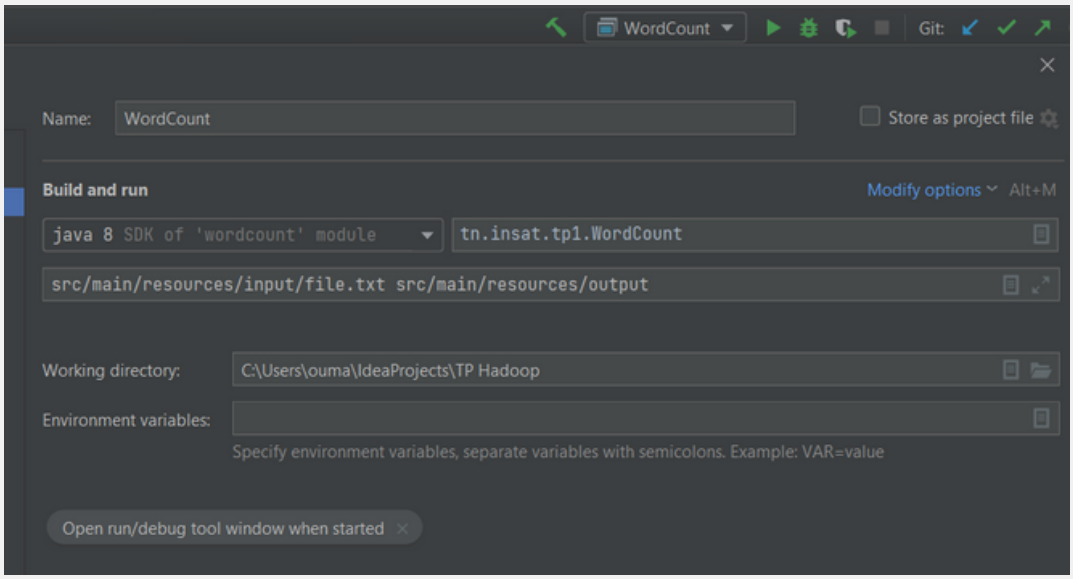
Word Count example

- Création du projet maven, utilisant le JDK 1.8
- Ajout des dépendences nécessaires pour Hadoop, HDFS et Map Reduce
- Création des packages et des classes nécessaire: TokenizerMapper(le Mapper), IntSumReducer (le reducer) et WordCount (main program)
- Création des ressources nécessaires

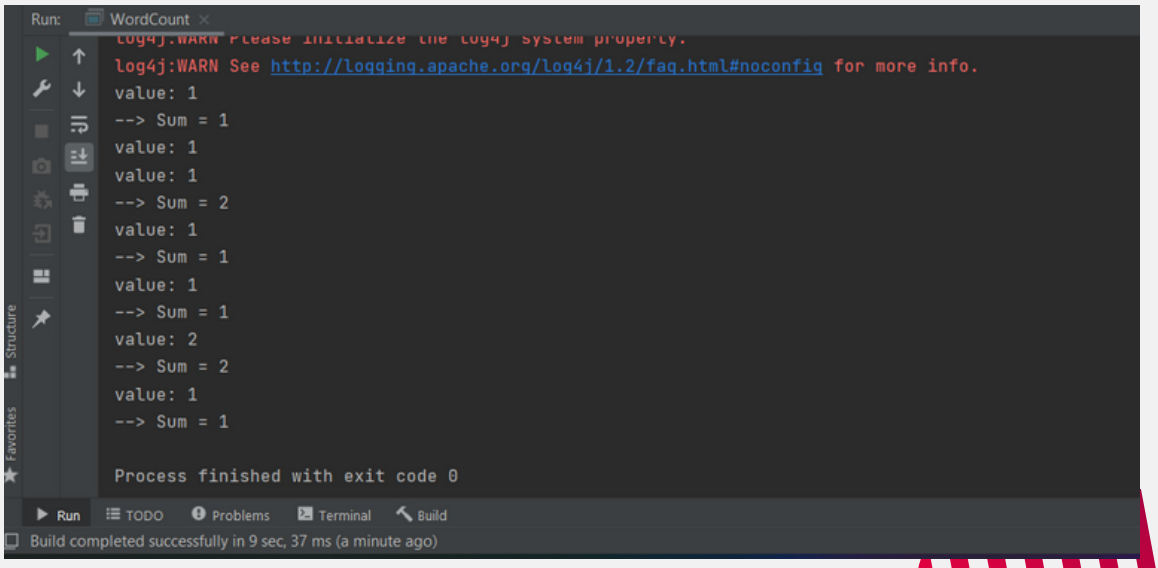


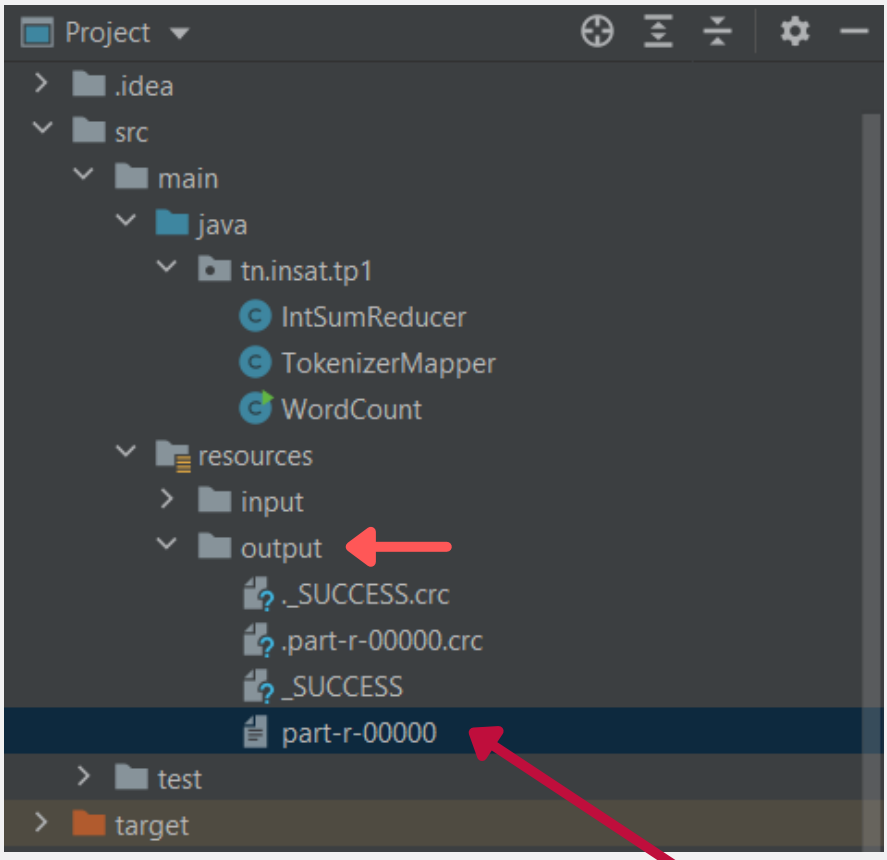


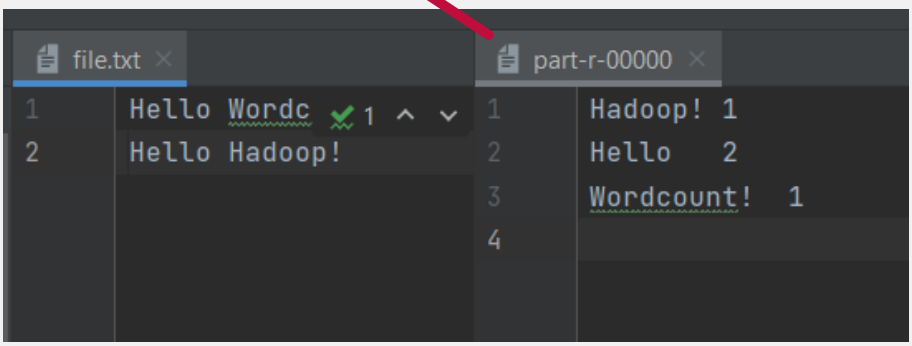
- Création de la configuration **application** adéquate à l'exemple **word count**



- Lancer le programme (d'où la création d'un répertoire **output** contenant le résultat)

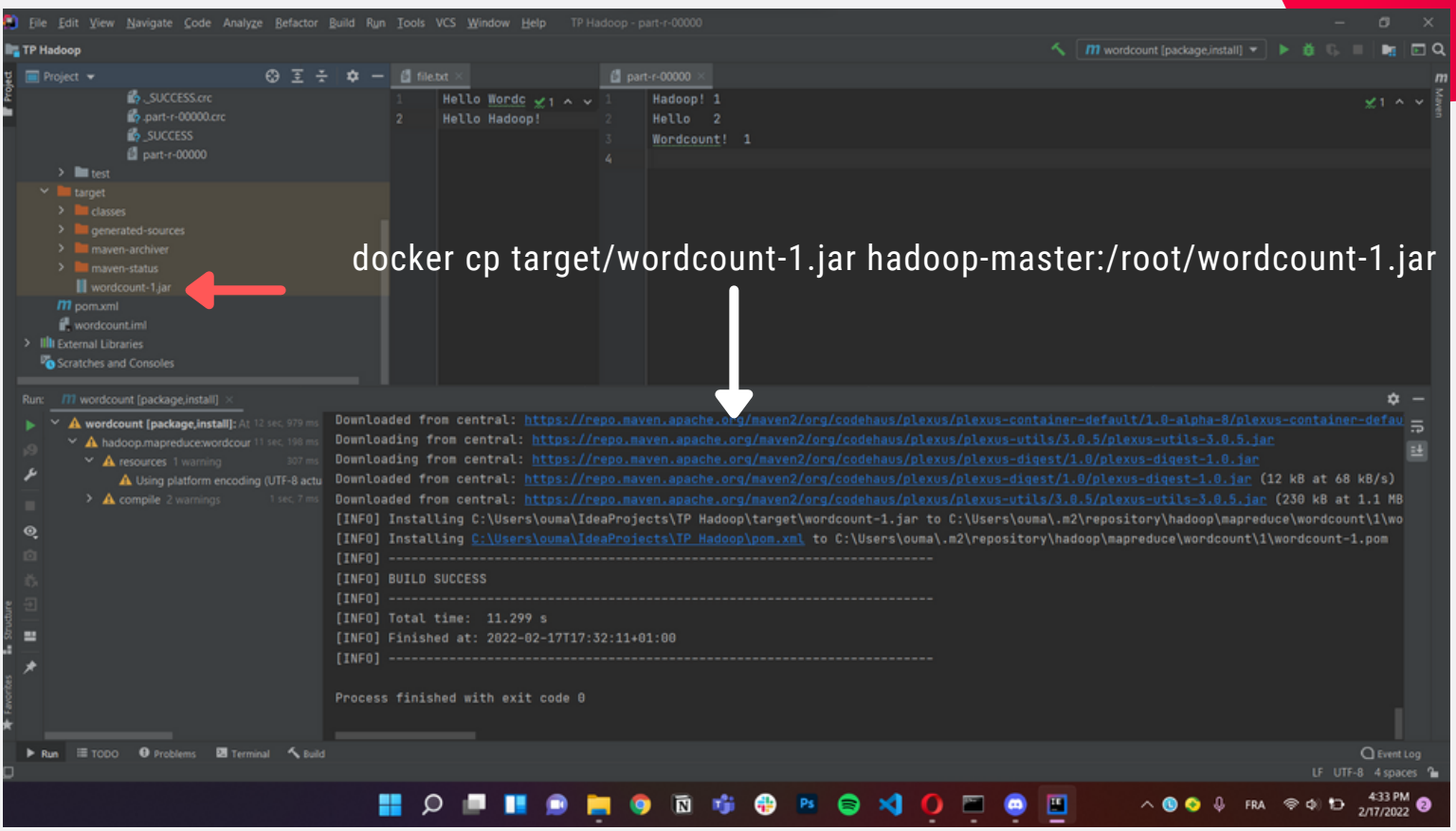






C'est le test de Map Reduce en local, on passe maintenant au test de Map Reduce sur le cluster

- Création de la configuration **Maven** adéquate à l'exemple **word count**



- lancer le job map reduce dans le contenaire master

```
root@hadoop-master:~# hadoop jar wordcount-1.jar tn.insat.tp1.WordCount input output
22/02/17 16:35:59 INFO client.RMProxy: Connecting to ResourceManager at hadoop-master/172.18.0.2:8032
22/02/17 16:35:59 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute
22/02/17 16:35:59 INFO input.FileInputFormat: Total input paths to process : 1
22/02/17 16:35:59 INFO mapreduce.JobSubmitter: number of splits:2
22/02/17 16:35:59 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1645104020056_0001
22/02/17 16:35:59 INFO impl.YarnClientImpl: Submitted application application_1645104020056_0001
22/02/17 16:35:59 INFO mapreduce.Job: The url to track the job: http://hadoop-master:8088/proxy/application_1645104020056_0001/
22/02/17 16:35:59 INFO mapreduce.Job: Running job: job_1645104020056_0001
22/02/17 16:36:04 INFO mapreduce.Job: Job job_1645104020056_0001 running in uber mode : false
22/02/17 16:36:04 INFO mapreduce.Job:  map 0% reduce 0%
22/02/17 16:36:15 INFO mapreduce.Job:  map 19% reduce 0%
22/02/17 16:36:30 INFO mapreduce.Job:  map 35% reduce 0%
22/02/17 16:36:45 INFO mapreduce.Job:  map 50% reduce 0%
22/02/17 16:37:03 INFO mapreduce.Job:  map 57% reduce 0%
22/02/17 16:37:04 INFO mapreduce.Job:  map 74% reduce 0%
22/02/17 16:37:15 INFO mapreduce.Job:  map 74% reduce 17%
22/02/17 16:37:18 INFO mapreduce.Job:  map 80% reduce 17%
22/02/17 16:37:33 INFO mapreduce.Job:  map 83% reduce 17%
22/02/17 16:37:42 INFO mapreduce.Job:  map 100% reduce 17%
22/02/17 16:37:44 INFO mapreduce.Job:  map 100% reduce 100%
22/02/17 16:37:44 INFO mapreduce.Job: Job job_1645104020056_0001 completed successfully
22/02/17 16:37:44 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=7684620
                FILE: Number of bytes written=9325480
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=211317260
                HDFS: Number of bytes written=499048
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=3
                Launched reduce tasks=1
                Data-local map tasks=3
                Total time spent by all maps in occupied slots (ms)=189065
                Total time spent by all reduces in occupied slots (ms)=36575
                Total time spent by all map tasks (ms)=189065
                Total time spent by all reduce tasks (ms)=36575
```

job running

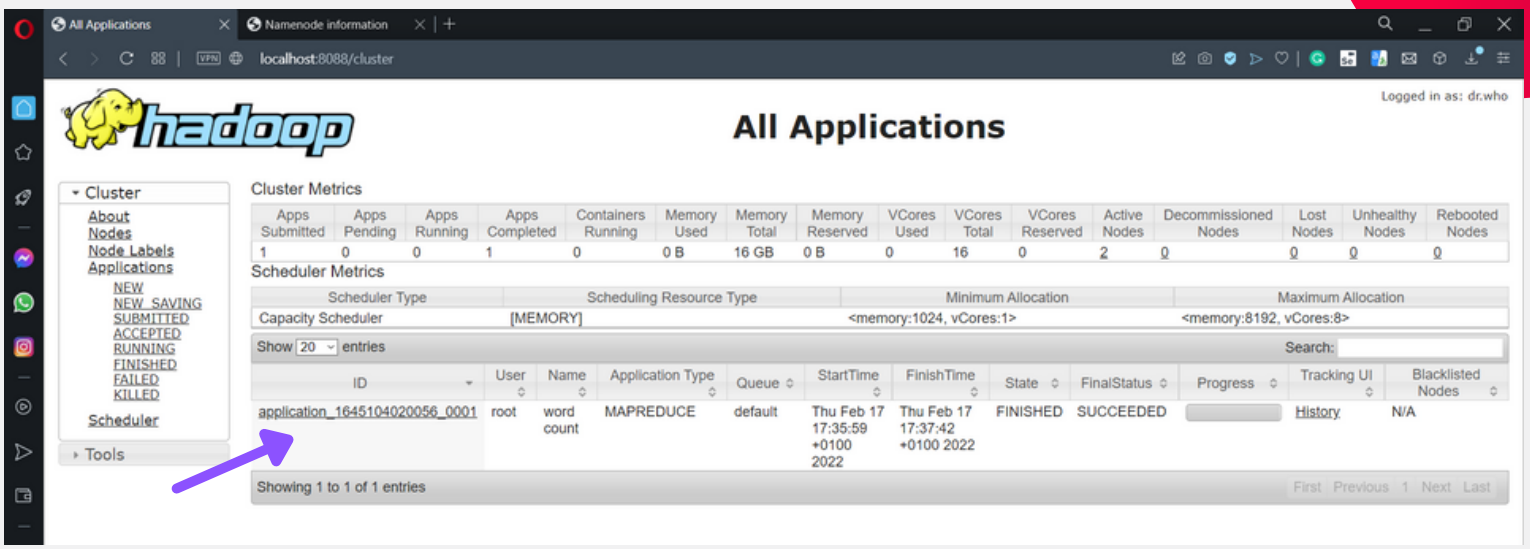
```
        Job Counters
                Killed map tasks=1
                Launched map tasks=3
                Launched reduce tasks=1
                Data-local map tasks=3
                Total time spent by all maps in occupied slots (ms)=189065
                Total time spent by all reduces in occupied slots (ms)=36575
                Total time spent by all map tasks (ms)=189065
                Total time spent by all reduce tasks (ms)=36575
                Total vcore-milliseconds taken by all map tasks=189065
                Total vcore-milliseconds taken by all reduce tasks=36575
                Total megabyte-milliseconds taken by all map tasks=193602560
                Total megabyte-milliseconds taken by all reduce tasks=37452800
        Map-Reduce Framework
                Map input records=4138476
                Map output records=27982895
                Map output bytes=323244504
                Map output materialized bytes=1289264
                Input split bytes=240
                Combine input records=28488079
                Combine output records=606926
                Reduce input groups=51053
                Reduce shuffle bytes=1289264
                Reduce input records=101742
                Reduce output records=51053
                Spilled Records=708668
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=525
                CPU time spent (ms)=162620
                Physical memory (bytes) snapshot=740212736
                Virtual memory (bytes) snapshot=5977616384
                Total committed heap usage (bytes)=553123840
        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=211317020
        File Output Format Counters
                Bytes Written=499048
root@hadoop-master:~#
```

- Afficher les dernières lignes du fichier généré output/part-r-00000 :

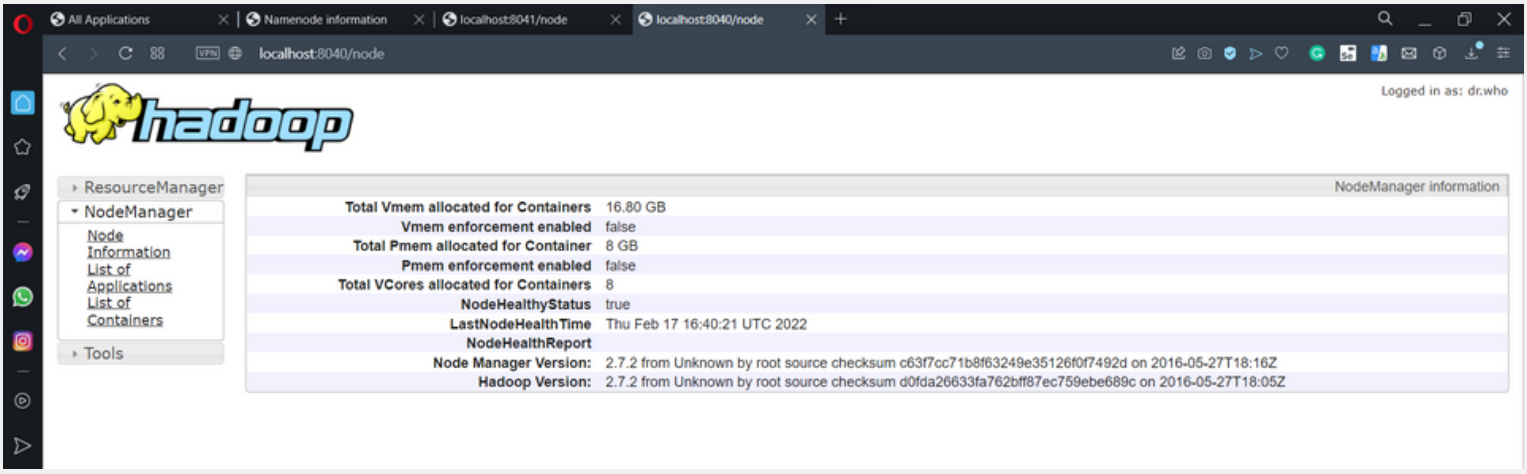
  hadoop fs -tail output/part-r-00000

```
root@hadoop-master: ~
Omaha    40209
Orlando 40195
Orleans 39914
Paso     39882
Paul     40160
Pet      229222
Petersburg      40093
Philadelphia    40748
Phoenix 40333
Pittsburgh      40358
Plano    40170
Portland        40065
Raleigh 40261
Reno     40254
Richmond        39983
Riverside       39963
Rochester       40455
Rouge    40387
Sacramento      40561
Saint    40160
San      200020
Santa    40306
Scottsdale      40173
Seattle 39866
Spokane 40222
Sporting        229932
Springs 40389
St.      80075
Stockton        39996
Supplies        229222
Tampa    40136
Toledo   40139
Toys     229964
Tucson   39870
Tulsa    40247
Vegas    80178
Video    230237
Virginia        40169
Visa     827221
Vista    40080
Washington      40503
Wayne    40439
Wichita 40422
Winston-Salem   40208
Women's 230050
Worth    40336
York     40364
and      229667
root@hadoop-master:~#
```

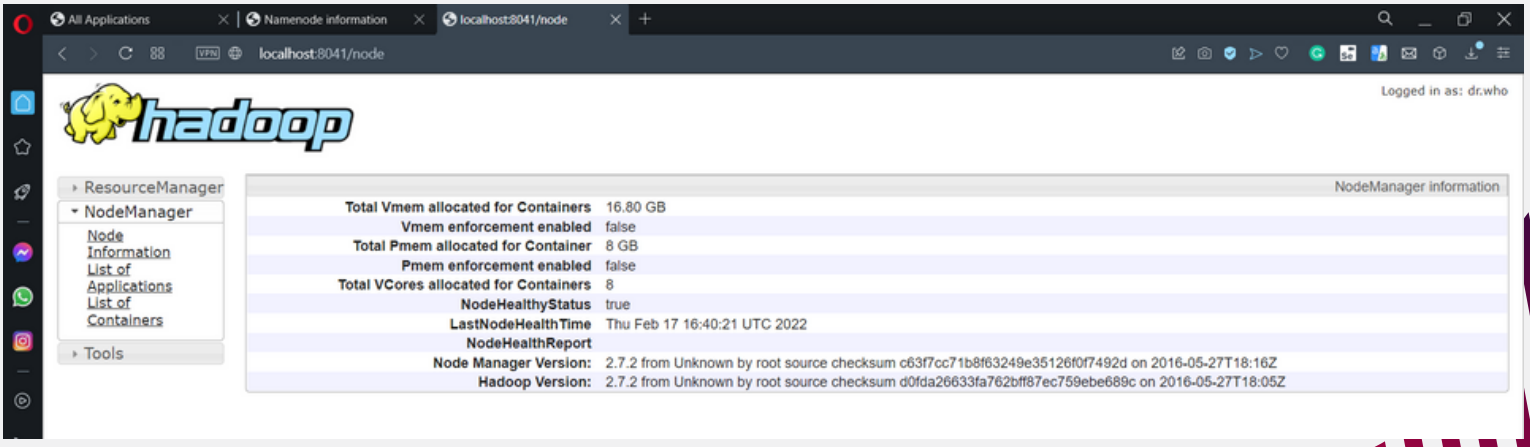
- Monitorer le Job Map Reduce en allant à la page **http://localhost:8088**



- Observer le comportement des noeuds esclaves, en allant à la page **http://localhost:8040** pour l'**esclave1**



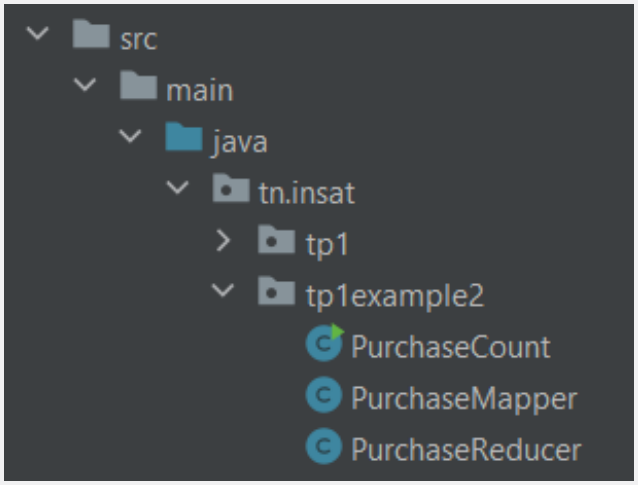
- Observer le comportement des noeuds esclaves, en allant à la page **http://localhost:8041** pour l'**esclave2**

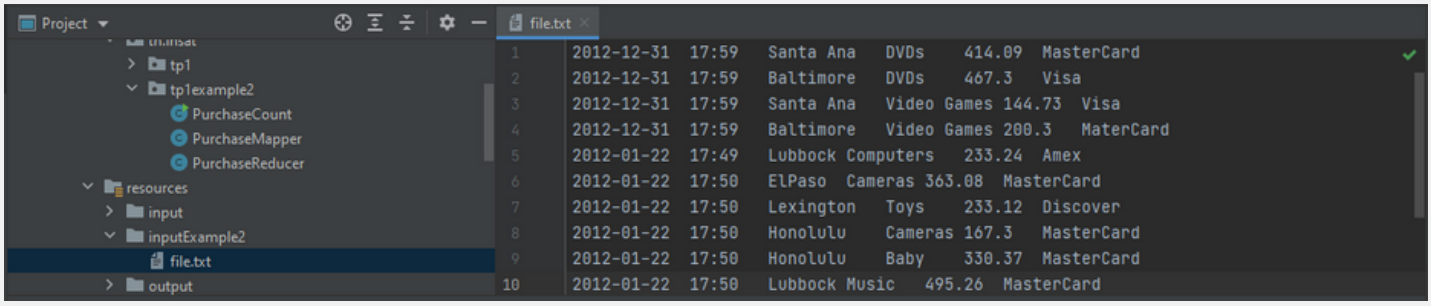


Map Reduce

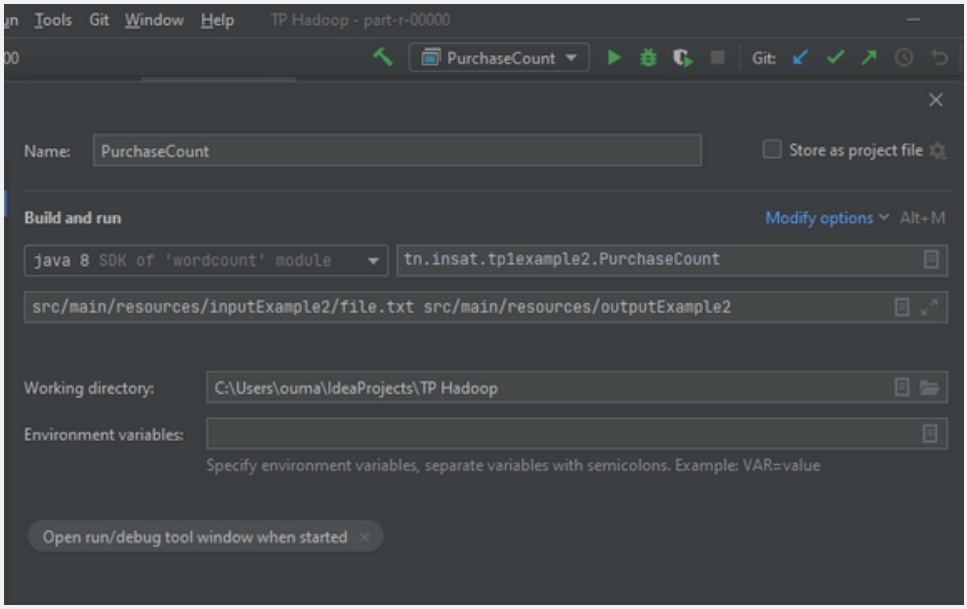
Purchase Count example

- Création des packages: tp1example2 et des classes nécessaire: PurchaseMapper(le Mapper), PurchaseReducer (le reducer) et PurchaseCount (main program)
- Création des ressources nécessaires

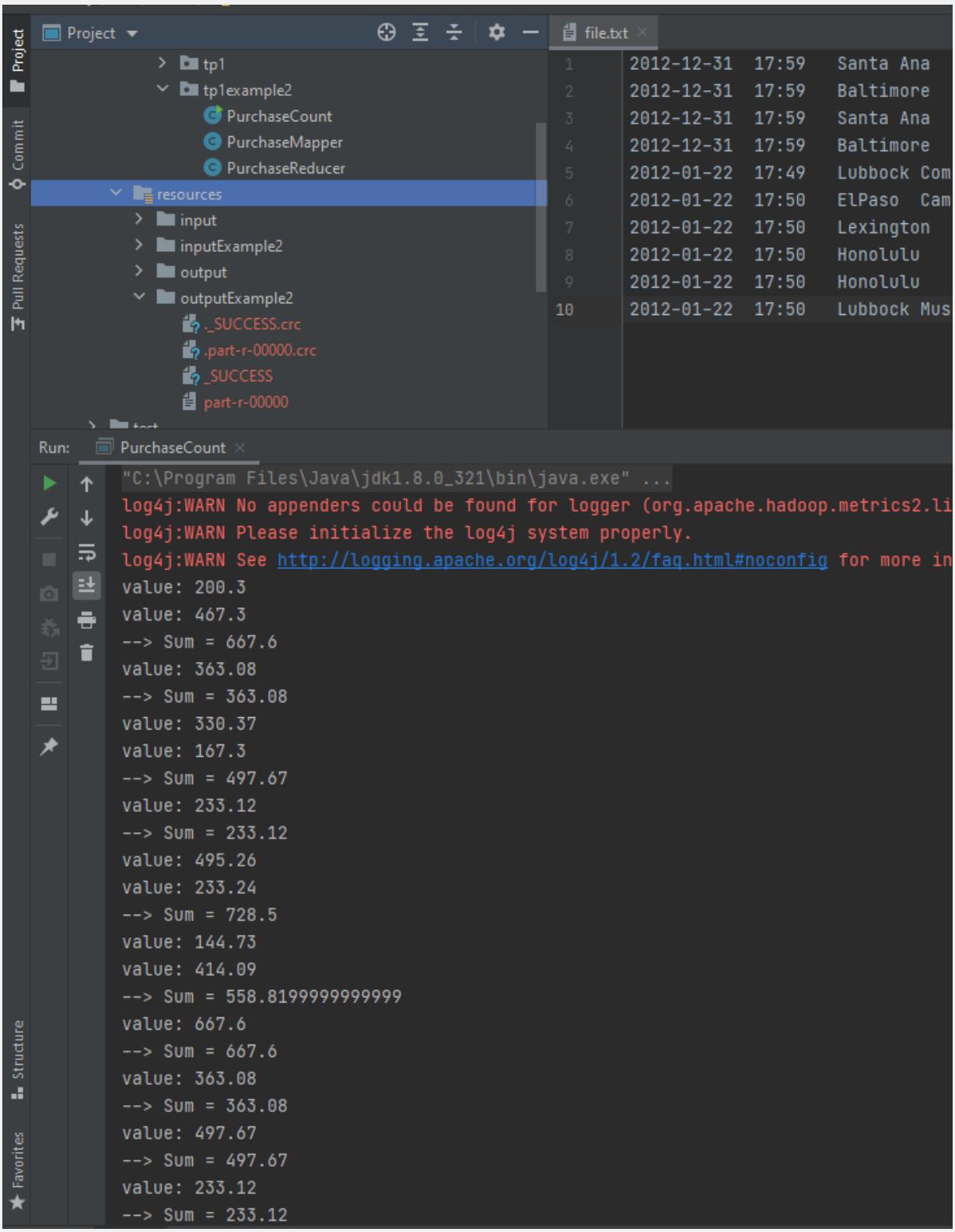


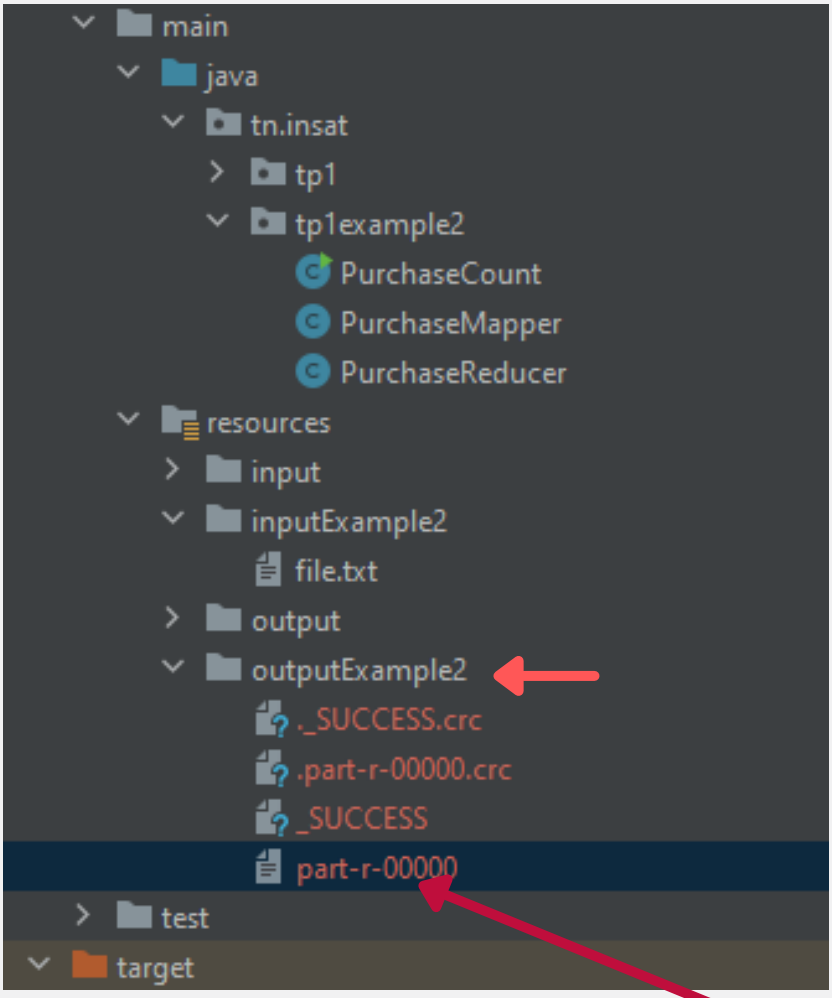


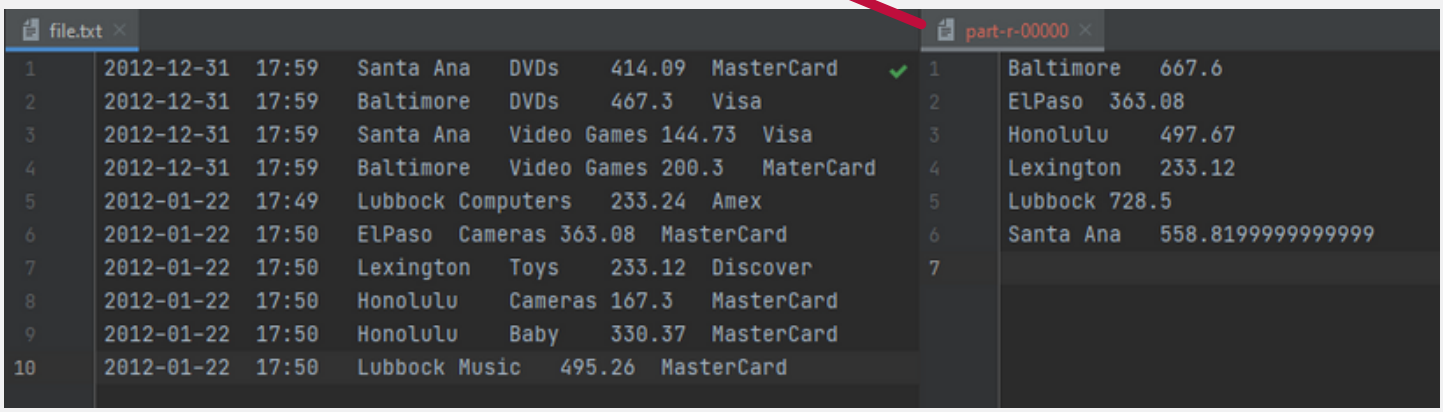
- Création de la configuration **application** adéquate à l'exemple **purchase count**



- Lancer le programme (d'où la création d'un répertoire **outputExample2** contenant le résultat)

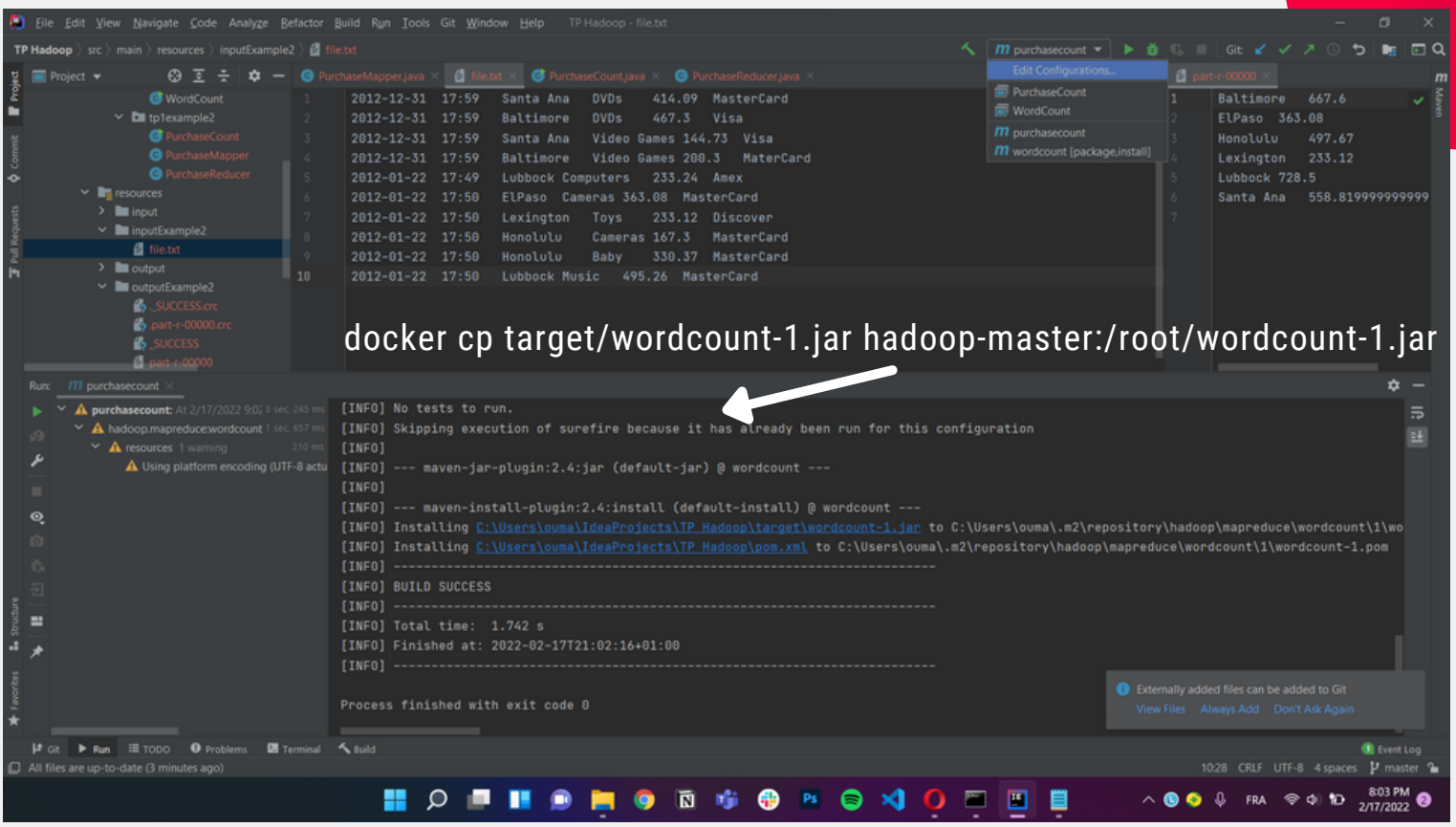




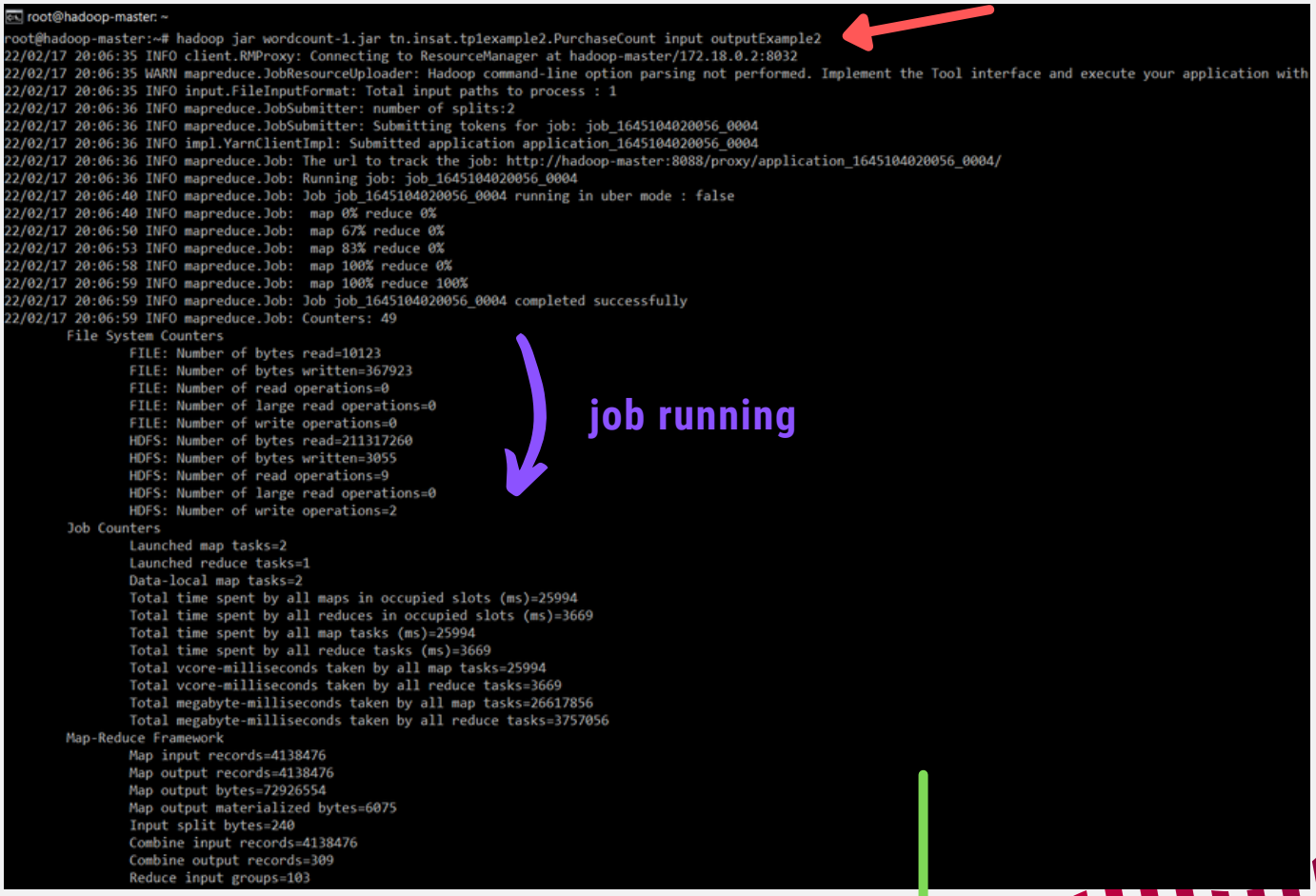


C'est le test de Map Reduce en local, on passe maintenant au test de Map Reduce sur le cluster

- Création de la configuration **Maven** adéquate à l'exemple **purchase count**



- lancer le job map reduce dans le contenaire master



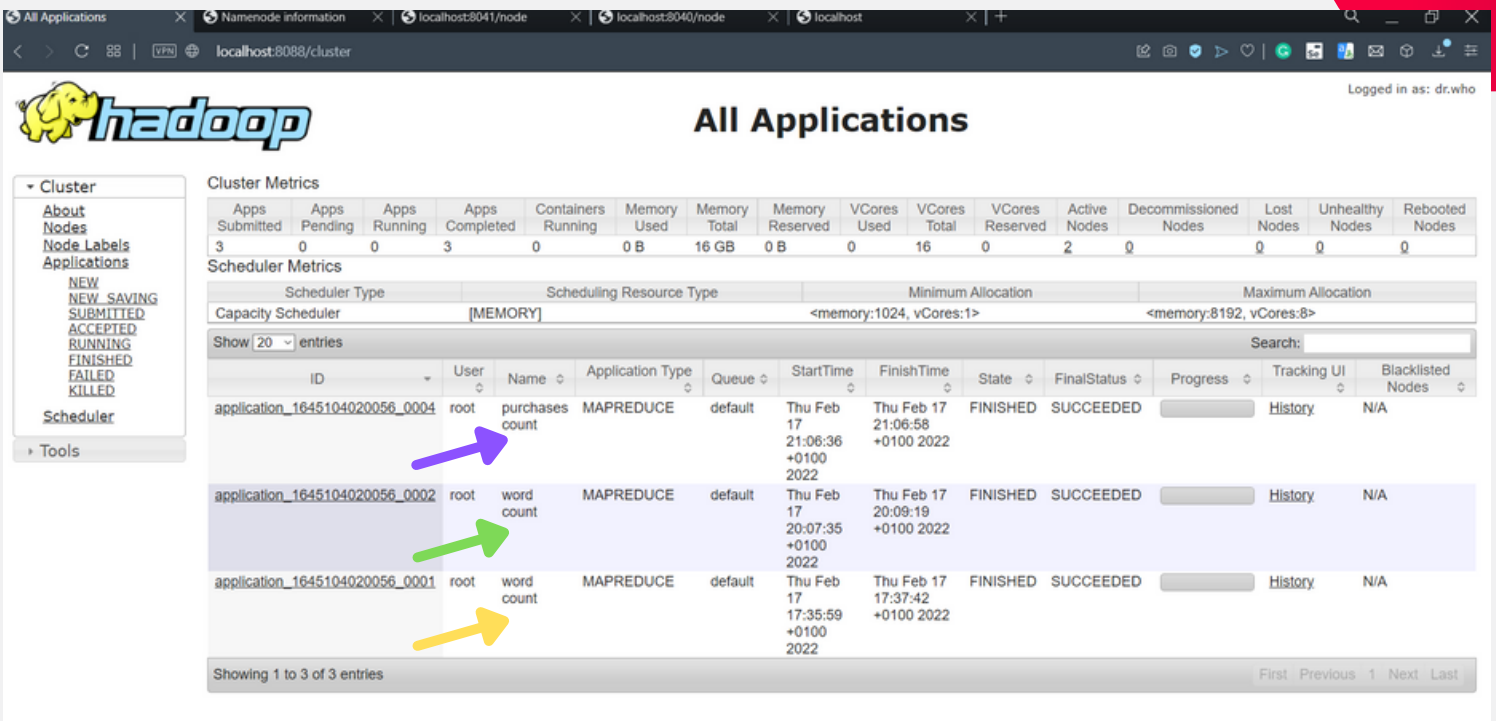
```
        Map-Reduce Framework
                Map input records=4138476
                Map output records=4138476
                Map output bytes=72926554
                Map output materialized bytes=6075
                Input split bytes=240
                Combine input records=4138476
                Combine output records=309
                Reduce input groups=103
                Reduce shuffle bytes=6075
                Reduce input records=309
                Reduce output records=103
                Spilled Records=824
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=295
                CPU time spent (ms)=28090
                Physical memory (bytes) snapshot=722550784
                Virtual memory (bytes) snapshot=5969301504
                Total committed heap usage (bytes)=565706752
        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=211317020
        File Output Format Counters
                Bytes Written=3055
```

- Afficher les dernières lignes du fichier généré output/part-r-00000 :

  hadoop fs -tail outputExample2/part-r-00000

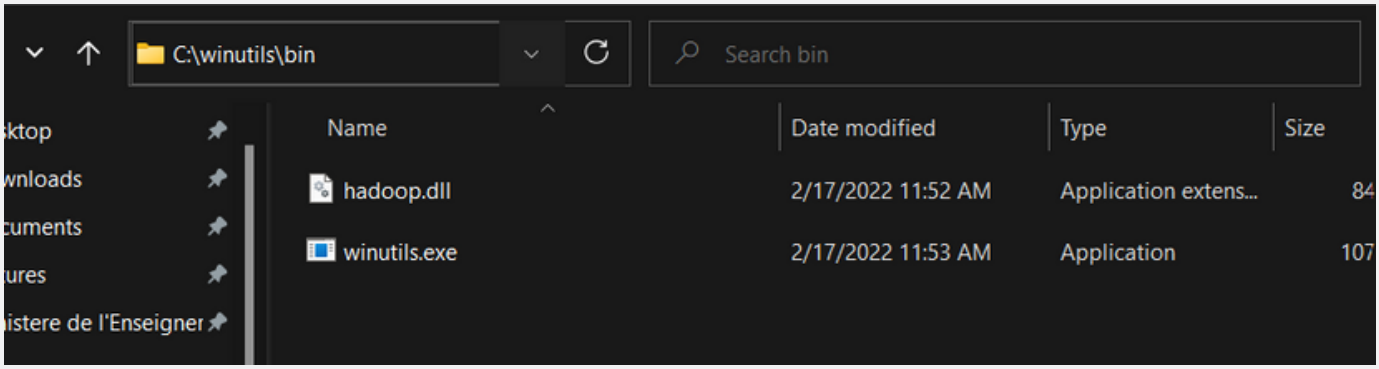
```
root@hadoop-master:~# hadoop fs -tail outputExample2/part-r-00000
98625000001E7
Omaha    1.0026642339999996E7
Orlando 1.0074922520000024E7
Philadelphia     1.0190080260000039E7
Phoenix 1.0079076700000018E7
Pittsburgh       1.0090124820000011E7
Plano    1.0046103609999988E7
Portland         1.000763576999994E7
Raleigh 1.006144253999997E7
Reno     1.0079955160000004E7
Richmond         9992941.589999985
Riverside        1.000669542E7
Rochester        1.0067606919999966E7
Sacramento       1.0123468179999985E7
Saint Paul       1.0057233570000034E7
San Antonio      1.0014441700000023E7
San Bernardino   9965152.039999895
San Diego        9966038.390000047
San Francisco    9995570.540000016
San Jose         9936721.409999996
Santa Ana        1.0050309929999996E7
Scottsdale       1.003792984999999E7
Seattle 9936267.370000027
Spokane 1.0083362979999978E7
St. Louis        1.0002105140000038E7
St. Petersburg   9986495.54000001
Stockton         1.0006412640000032E7
Tampa    1.0106428549999947E7
Toledo   1.0020768880000012E7
Tucson   9998252.469999975
Tulsa    1.0064955900000023E7
Virginia Beach   1.0086553500000041E7
Washington       1.013936339000001E7
Wichita 1.0083643210000023E7
Winston-Salem    1.0044011829999976E7
root@hadoop-master:~# hadoop fs -tail outputExample2/part-r-00000
98625000001E7
Omaha    1.0026642339999996E7
Orlando 1.0074922520000024E7
Philadelphia     1.0190080260000039E7
Phoenix 1.0079076700000018E7
Pittsburgh       1.0090124820000011E7
Plano    1.0046103609999988E7
Portland         1.000763576999994E7
Raleigh 1.006144253999997E7
```

- Monitorer le Job Map Reduce en allant à la page **http://localhost:8088**

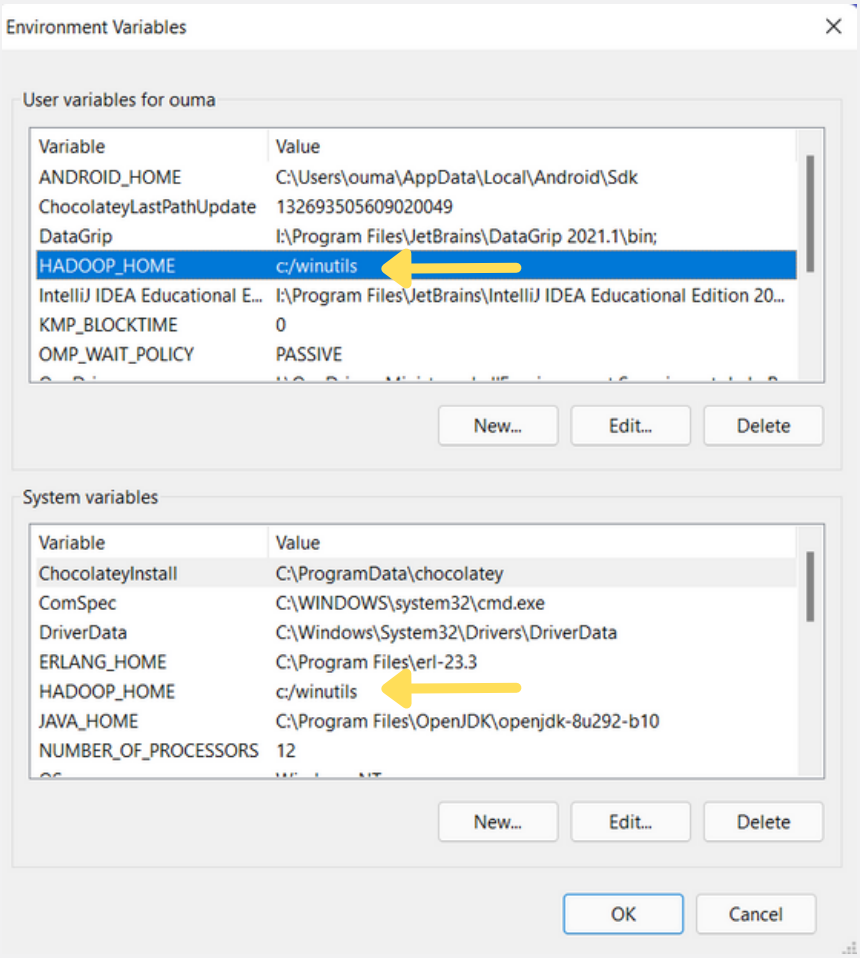


Remarque:

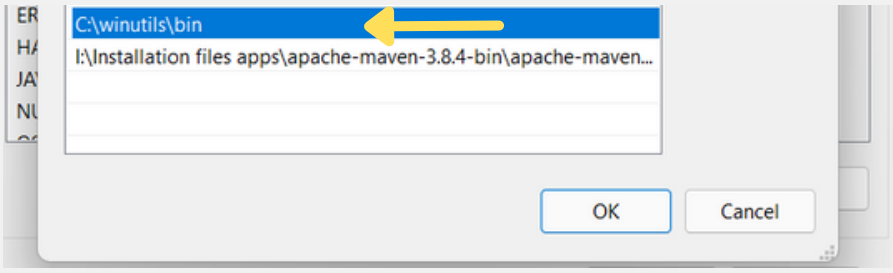
- Il faut ajouter les deux fichier **hadoop.dll** et **winutils.exe** dans **C:\winutils\bin**
- Pour obtenir les fichier consulter ce répo: **https://github.com/cdarlint/winutils/tree/master/hadoop-2.7.2/bin**



Ajouter la variable **HADOOP_HOME** avec le path: **c:/winutils**



Ajouter le path **C:\winutils\bin** de bin dans l'environement



Le code est dans le répo suivant:

https://github.com/oumaima-kboubi/Big-Data-Hadoop-FirstSteps

THE END