

Compte Rendu TP1



GL4

Hadoop et Map Reduce

Binômes

- Oumaima KBOUBI
- Rami Kammoun

Installation



Après 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 conteneurs, on crée ces 3 conteneurs et on les lance

```
Command Prompt

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

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

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

C:\Users\ouma>
```

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

```
C:\Users\ouma>docker run -itd --net=hadoop -p 8041:8042 --name hadoop-slave2 --hostname hadoop-slave2 liliasfaxi/spark-hadoop:hv-2.7.2
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
root@hadoop-master:~# ls
hdfs purchases.txt purchases2.txt run-wordcount.sh start-hadoop.sh start-kafka-zookeeper.sh
root@hadoop-master:~# hadoop fs -put purchases.txt input
root@hadoop-master:~# hadoop fs -ls input
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
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

The screenshot shows the 'Namenode Information' web interface. The browser address bar displays 'localhost:50070/dfshealth.html#tab-overview'. The interface has a green header with tabs: 'Hadoop', 'Overview', 'Datanodes', 'Datanode Volume Failures', 'Snapshot', 'Startup Progress', and 'Utilities'. The 'Overview' tab is selected, showing 'Overview 'hadoop-master:9000' (active)'. Below this, there is a table with the following data:

Started:	Thu Feb 17 13:20:06 UTC 2022
Version:	2.7.2, rUnknown
Compiled:	2016-05-27T18:05Z by root from Unknown
Cluster ID:	CID-b721bea8-93cb-45f0-9023-dff705808b00
Block Pool ID:	BP-195763961-172.17.0.3-1550840521902

Below the table is a 'Summary' section with the following text:

Security is off.
Safemode is off.
5 files and directories, 2 blocks = 7 total filesystem object(s).
Heap Memory used 75.01 MB of 222.5 MB Heap Memory. Max Heap Memory is 889 MB.
Non Heap Memory used 40.8 MB of 41.44 MB Committed Non Heap Memory. Max Non Heap Memory is -1 B.

Below the summary is another table:

Configured Capacity:	501.96 GB
DFS Used:	406.24 MB (0.08%)
Non DFS Used:	33.41 GB

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.

The screenshot shows the 'All Applications' web interface. The browser address bar displays 'localhost:8088/cluster'. The interface has a yellow elephant logo and the word 'hadoop' in blue. The title is 'All Applications'. On the left, there is a sidebar with a 'Cluster' section containing links: 'About', 'Nodes', 'Node Labels', 'Applications', 'NEW', 'NEW SAVING', 'SUBMITTED', 'ACCEPTED', 'RUNNING', 'FINISHED', 'FAILED', 'KILLED', 'Scheduler', and 'Tools'. The main content area shows 'Cluster Metrics' with a table:

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
0	0	0	0	0	0 B	16 GB	0 B	0	16	0	2	0	0	0	0

Below the table is a 'Scheduler Metrics' section with a table:

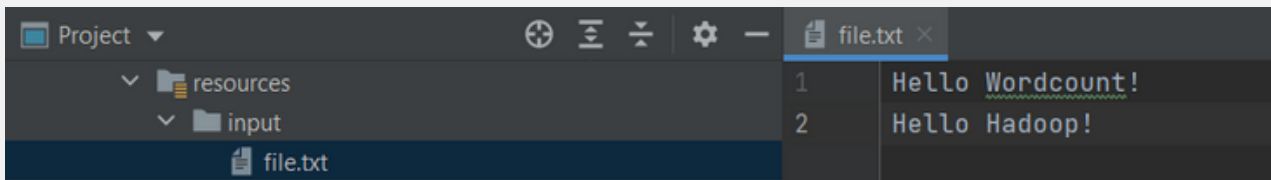
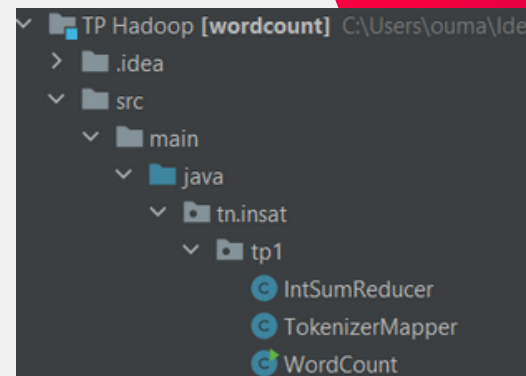
Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[MEMORY]	<memory:1024, vCores:1>	<memory:8192, vCores:32>

Below the scheduler metrics is a table with columns: ID, User, Name, Application Type, Queue, StartTime, FinishTime, State, FinalStatus, Progress, Tracking UI, and Blacklisted Nodes. The table is empty, showing 'No data available in table'.

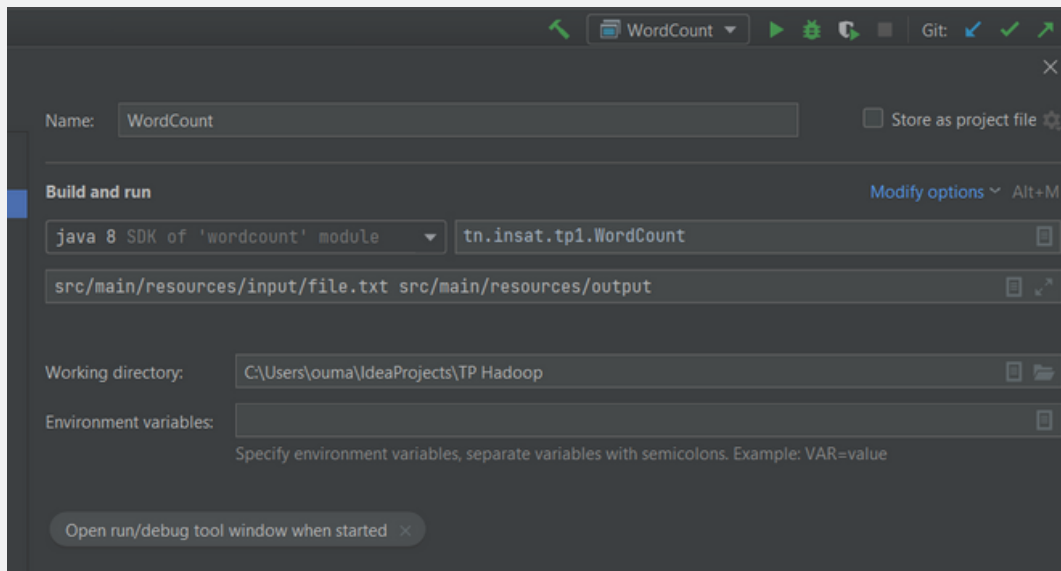
Map Reduce

Word Count example

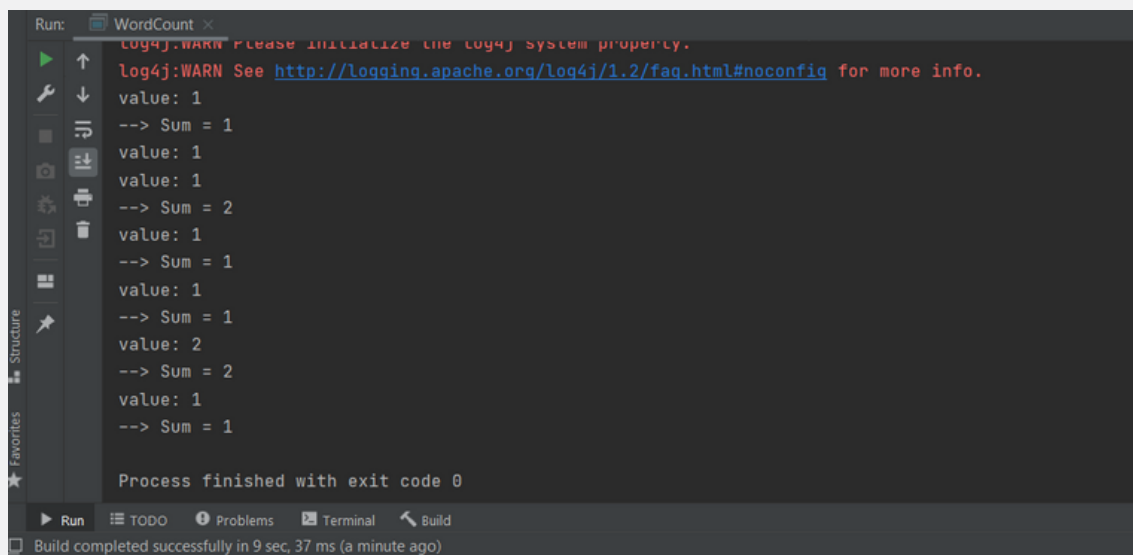
- Création du projet maven, utilisant le JDK 1.8
- Ajout des dépendances nécessaires pour Hadoop, HDFS et Map Reduce
- Création des packages et des classes nécessaire: TokenizerMapper(le Mapper), IntSumReducer (le reducer) et WordCount (main program)
- Création des ressources nécessaires

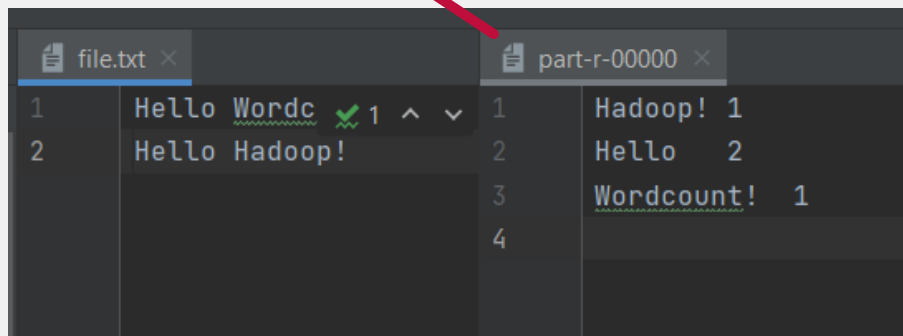
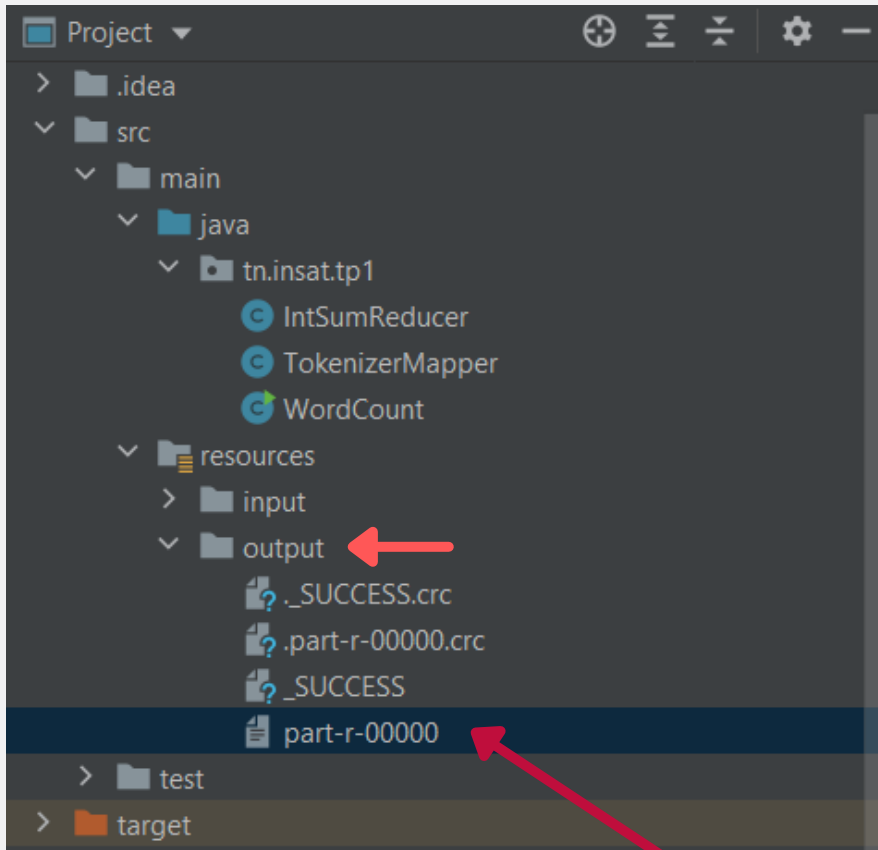


- Création de la configuration **application** adéquate à l'exemple **word count**



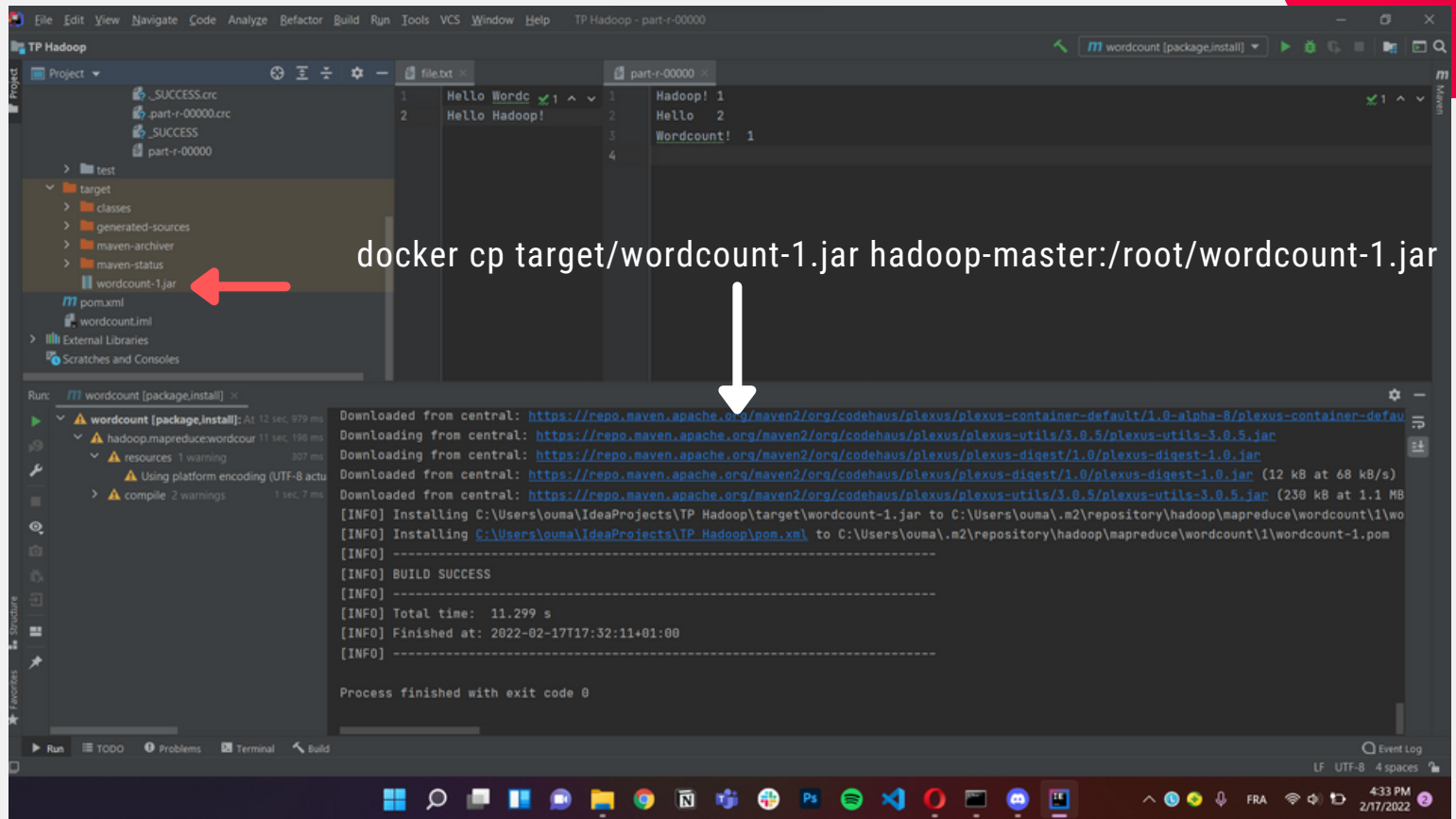
- Lancer le programme (d'où la création d'un répertoire **output** contenant le résultat)



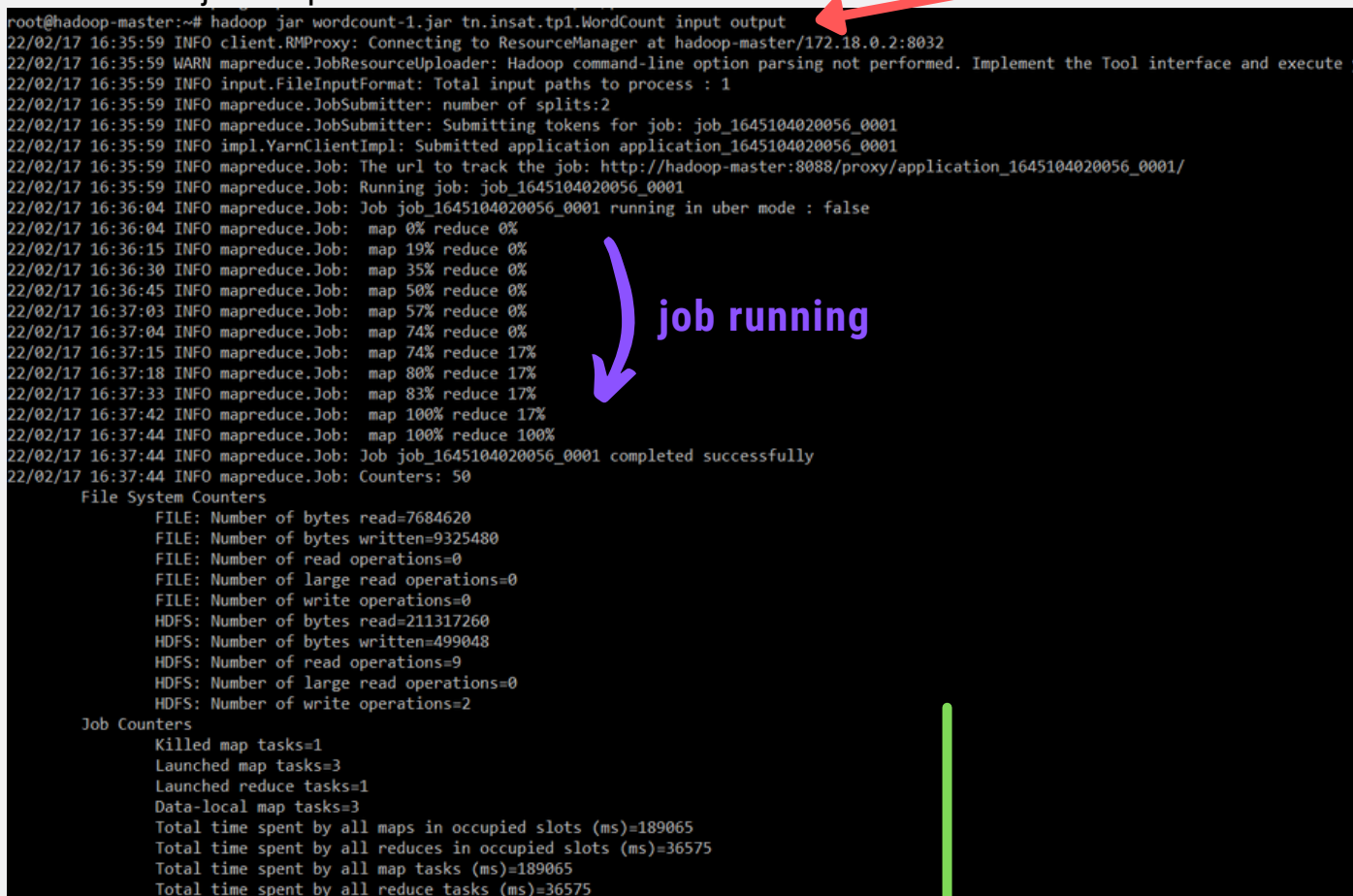


C'est le test de Map Reduce en local, on passe maintenant au test de Map Reduce sur le cluster

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



- lancer le job map reduce dans le conteneur master



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>**

All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
1	0	0	1	0	0 B	16 GB	0 B	0	16	0	2	0	0	0	0

Scheduler Metrics

Scheduler Type		Scheduling Resource Type		Minimum Allocation		Maximum Allocation	
Capacity Scheduler		[MEMORY]		<memory:1024, vCores:1>		<memory:8192, vCores:8>	

Show 20 entries

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tracking UI	Blacklisted Nodes
application_1645104020056_0001	root	word count	MAPREDUCE	default	Thu Feb 17 17:35:59 +0100 2022	Thu Feb 17 17:37:42 +0100 2022	FINISHED	SUCCEEDED		History	N/A

Showing 1 to 1 of 1 entries

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

NodeManager information

Total Vmem allocated for Containers	16.80 GB
Vmem enforcement enabled	false
Total Pmem allocated for Container	8 GB
Pmem enforcement enabled	false
Total VCores allocated for Containers	8
NodeHealthyStatus	true
LastNodeHealthTime	Thu Feb 17 16:40:21 UTC 2022
NodeHealthReport	
Node Manager Version:	2.7.2 from Unknown by root source checksum c63f7cc71b8f63249e35126f0f7492d on 2016-05-27T18:16Z
Hadoop Version:	2.7.2 from Unknown by root source checksum d0fda26633fa762bfff87ec759ebe689c on 2016-05-27T18:05Z

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

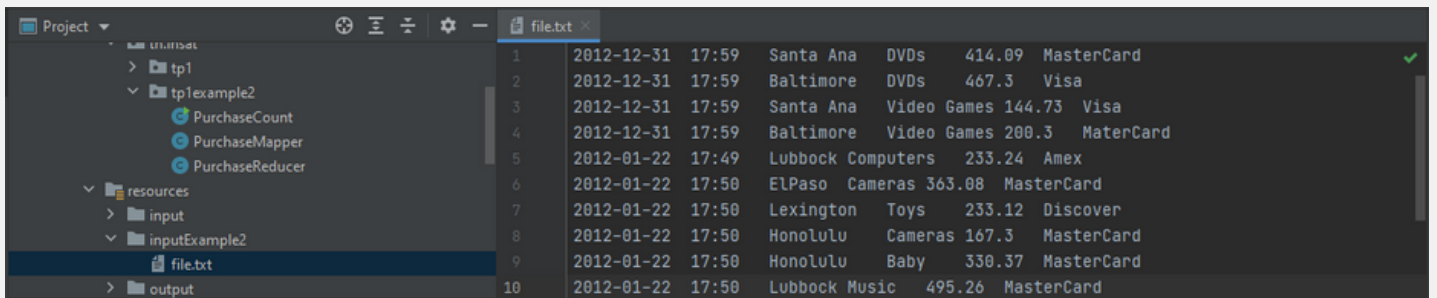
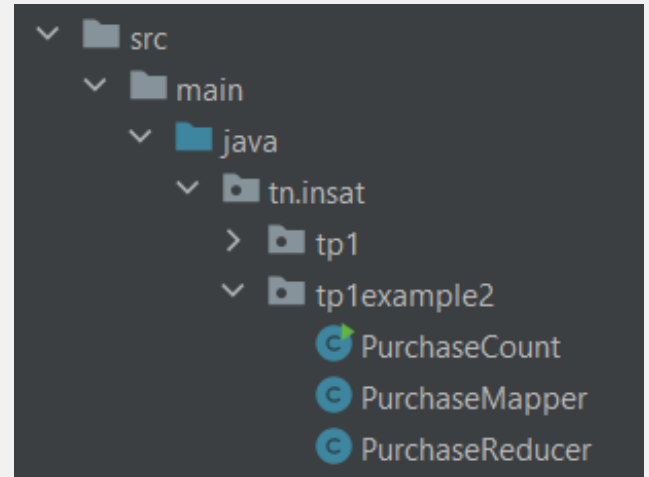
NodeManager information

Total Vmem allocated for Containers	16.80 GB
Vmem enforcement enabled	false
Total Pmem allocated for Container	8 GB
Pmem enforcement enabled	false
Total VCores allocated for Containers	8
NodeHealthyStatus	true
LastNodeHealthTime	Thu Feb 17 16:40:21 UTC 2022
NodeHealthReport	
Node Manager Version:	2.7.2 from Unknown by root source checksum c63f7cc71b8f63249e35126f0f7492d on 2016-05-27T18:16Z
Hadoop Version:	2.7.2 from Unknown by root source checksum d0fda26633fa762bfff87ec759ebe689c on 2016-05-27T18:05Z

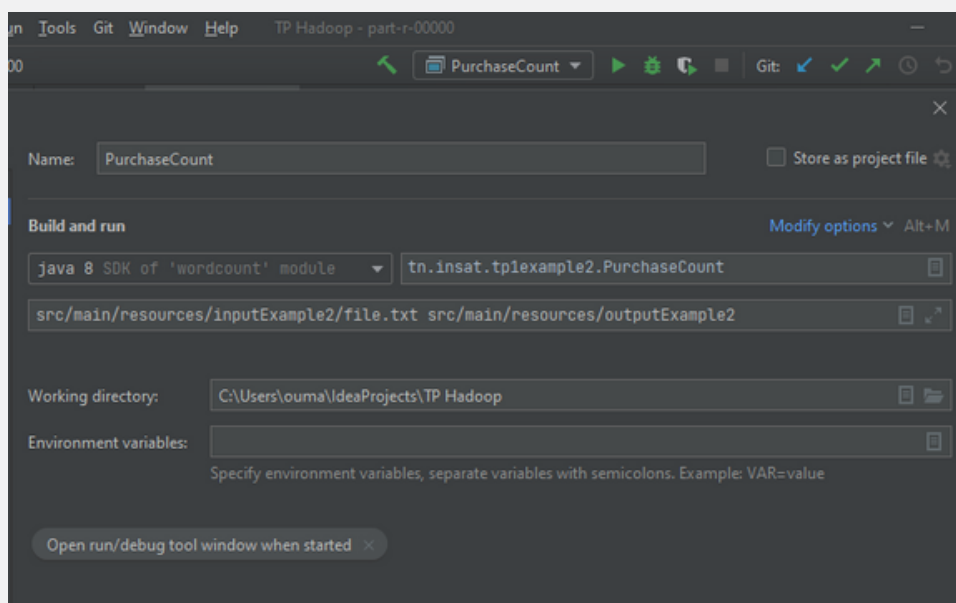
Map Reduce

Purchase Count example

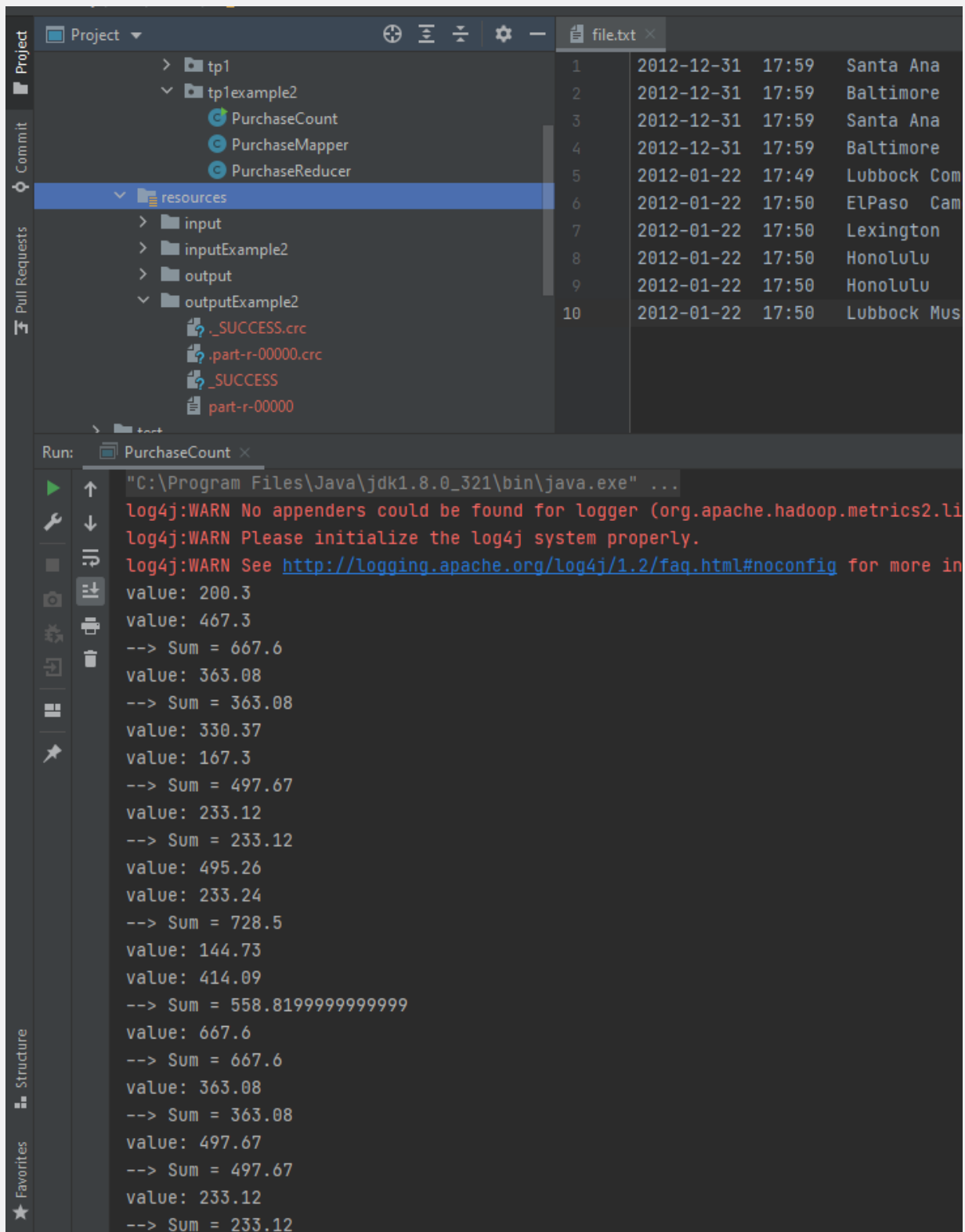
- Création des packages: tp1example2 et des classes nécessaire: PurchaseMapper(le Mapper), PurchaseReducer (le reducer) et PurchaseCount (main program)
- Création des ressources nécessaires



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

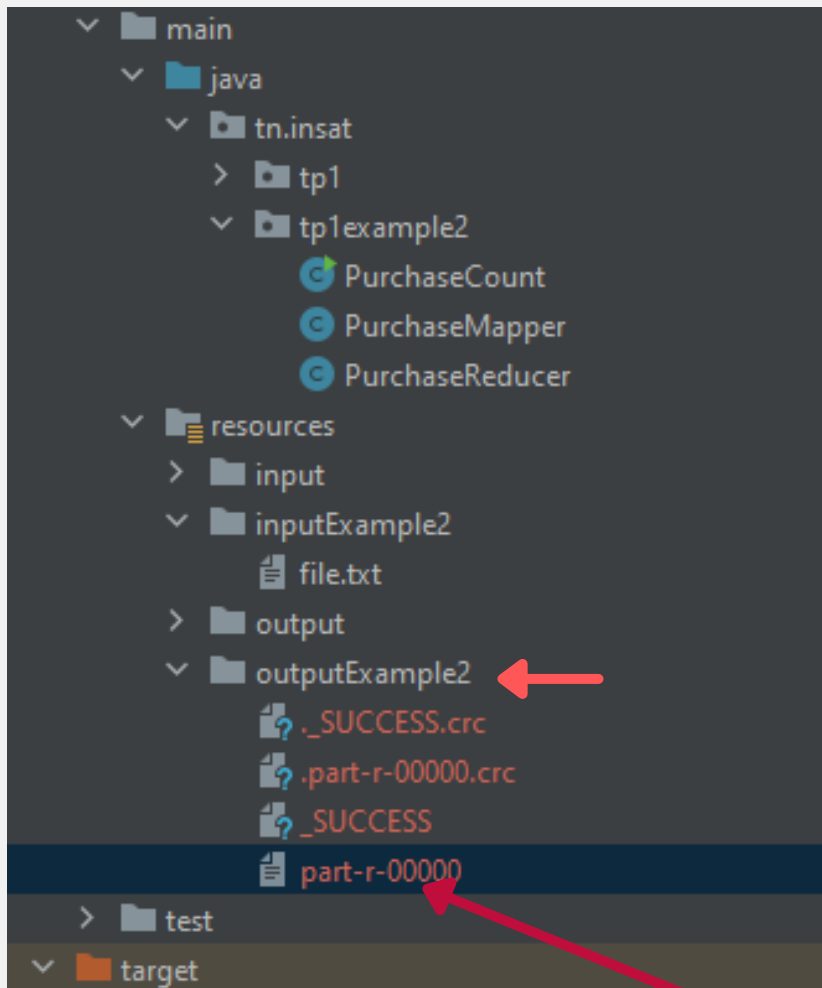


- Lancer le programme (d'où la création d'un répertoire **outputExample2** contenant le résultat)



The screenshot shows an IDE with a project structure on the left and a console output at the bottom. The project structure includes a folder named **tp1example2** containing **PurchaseCount**, **PurchaseMapper**, and **PurchaseReducer**. Below this is a **resources** folder containing **input**, **inputExample2**, **output**, and **outputExample2**. The **outputExample2** folder contains files like **._SUCCESS.crc**, **.part-r-00000.crc**, **._SUCCESS**, and **part-r-00000**. The console output shows the execution of **PurchaseCount** with various log messages and numerical values.

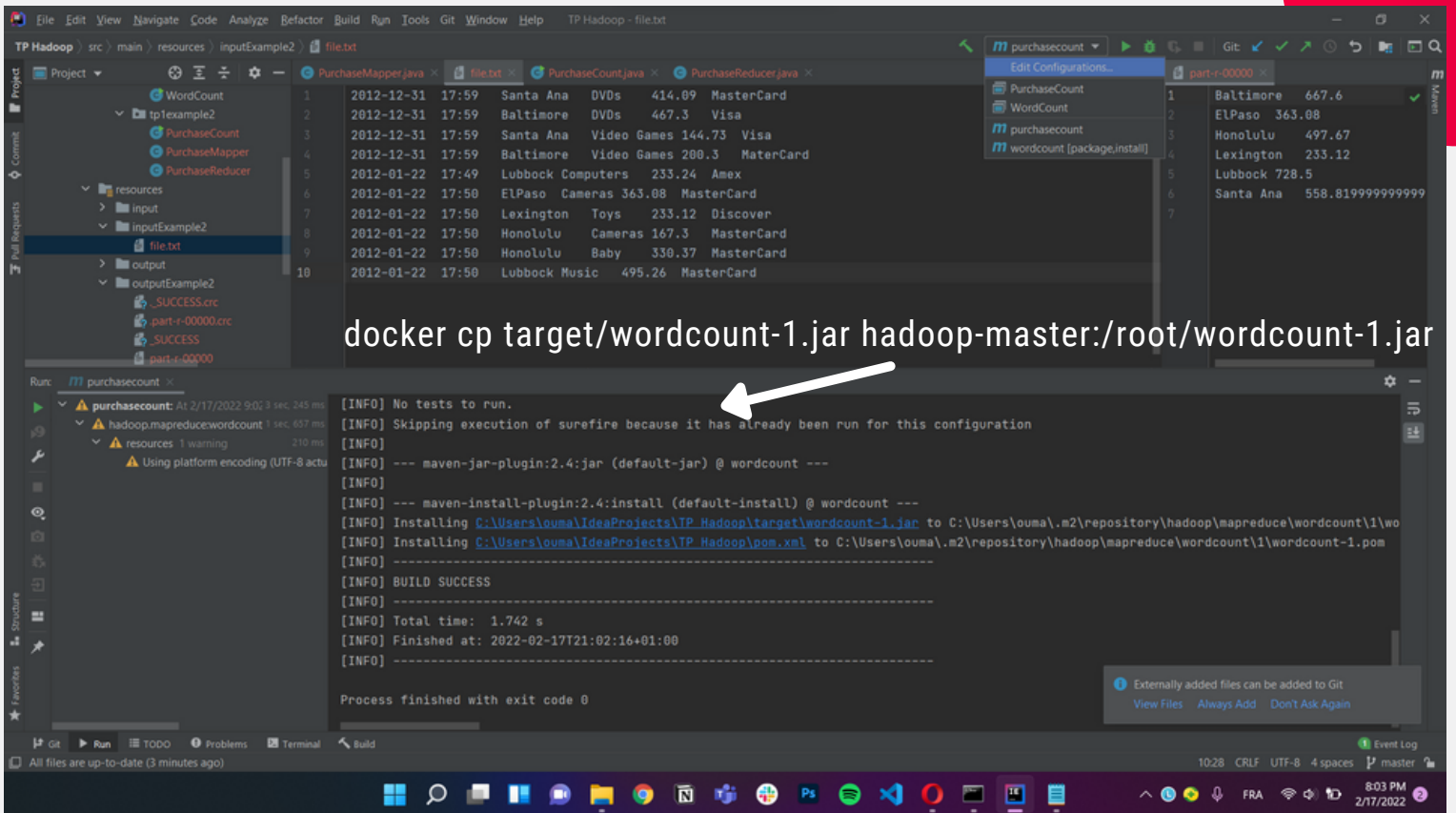
```
"C:\Program Files\Java\jdk1.8.0_321\bin\java.exe" ...  
log4j:WARN No appenders could be found for logger (org.apache.hadoop.metrics2.li  
log4j:WARN Please initialize the log4j system properly.  
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in  
value: 200.3  
value: 467.3  
--> Sum = 667.6  
value: 363.08  
--> Sum = 363.08  
value: 330.37  
value: 167.3  
--> Sum = 497.67  
value: 233.12  
--> Sum = 233.12  
value: 495.26  
value: 233.24  
--> Sum = 728.5  
value: 144.73  
value: 414.09  
--> Sum = 558.8199999999999  
value: 667.6  
--> Sum = 667.6  
value: 363.08  
--> Sum = 363.08  
value: 497.67  
--> Sum = 497.67  
value: 233.12  
--> Sum = 233.12
```



file.txt							part-r-00000		
1	2012-12-31	17:59	Santa Ana	DVDs	414.09	MasterCard	✓	1	Baltimore 667.6
2	2012-12-31	17:59	Baltimore	DVDs	467.3	Visa		2	ElPaso 363.08
3	2012-12-31	17:59	Santa Ana	Video Games	144.73	Visa		3	Honolulu 497.67
4	2012-12-31	17:59	Baltimore	Video Games	200.3	MasterCard		4	Lexington 233.12
5	2012-01-22	17:49	Lubbock	Computers	233.24	Amex		5	Lubbock 728.5
6	2012-01-22	17:50	ElPaso	Cameras	363.08	MasterCard		6	Santa Ana 558.8199999999999
7	2012-01-22	17:50	Lexington	Toys	233.12	Discover		7	
8	2012-01-22	17:50	Honolulu	Cameras	167.3	MasterCard			
9	2012-01-22	17:50	Honolulu	Baby	330.37	MasterCard			
10	2012-01-22	17:50	Lubbock	Music	495.26	MasterCard			

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**



docker cp target/wordcount-1.jar hadoop-master:/root/wordcount-1.jar

- lancer le job map reduce dans le conteneur master



job running

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

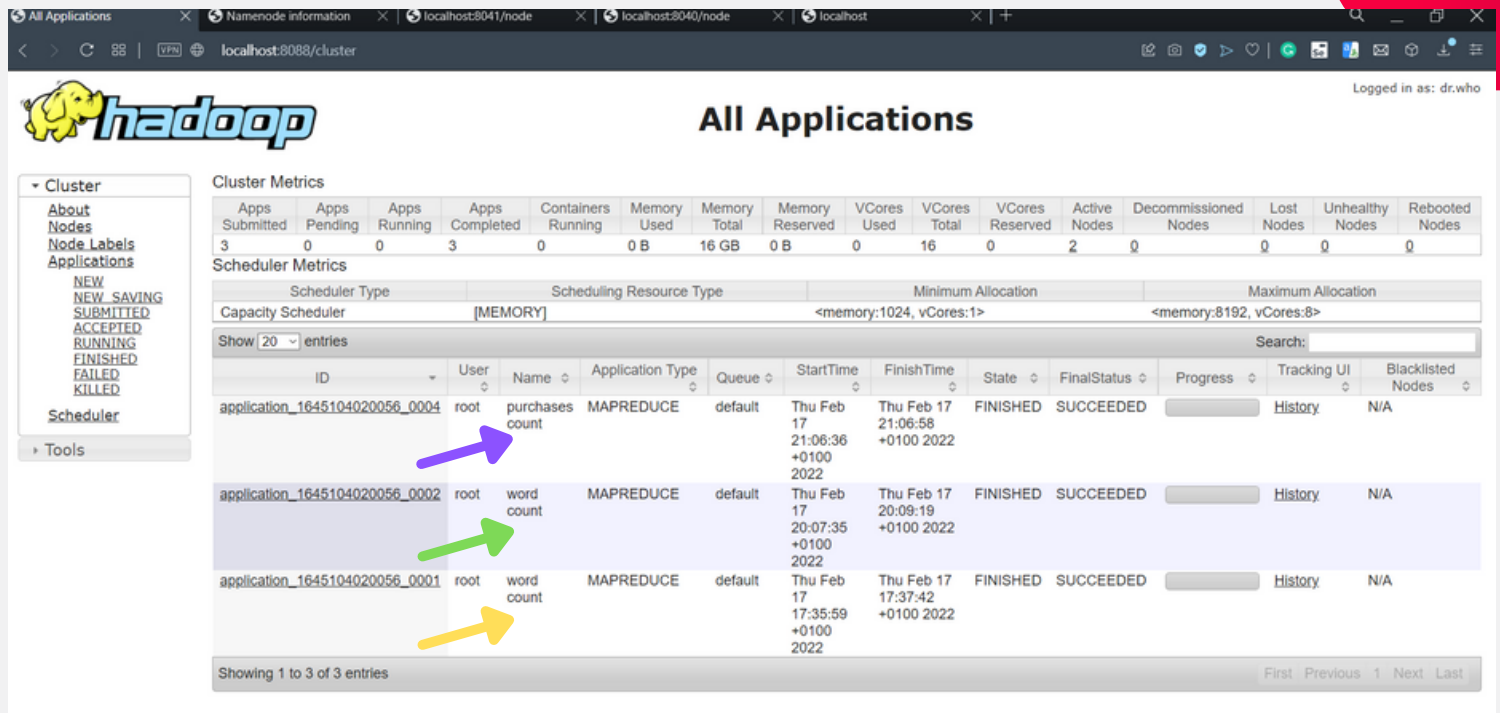
Bytes Written=3055

- 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
986250000001E7
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>**



Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
3	0	0	3	0	0 B	16 GB	0 B	0	16	0	2	0	0	0	0

Scheduler Metrics

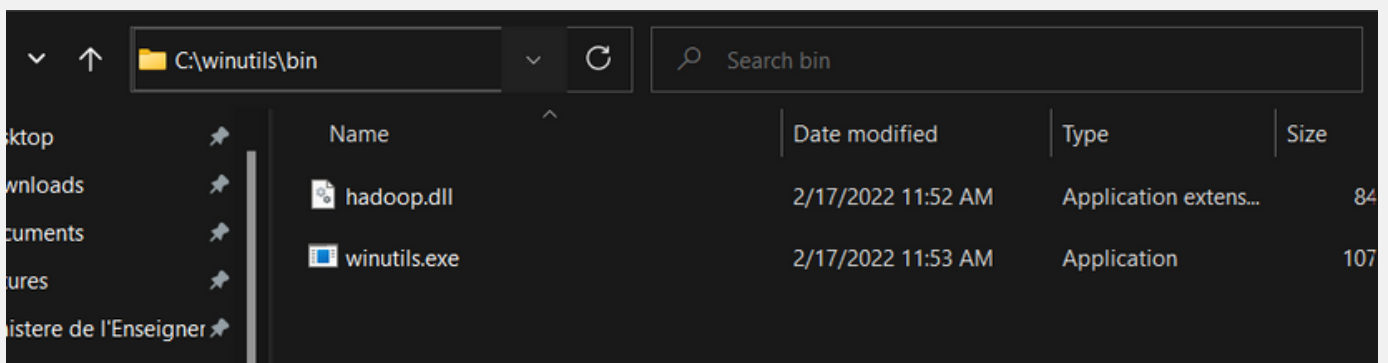
Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[MEMORY]	<memory:1024, vCores:1>	<memory:8192, vCores:8>

Showing 1 to 3 of 3 entries

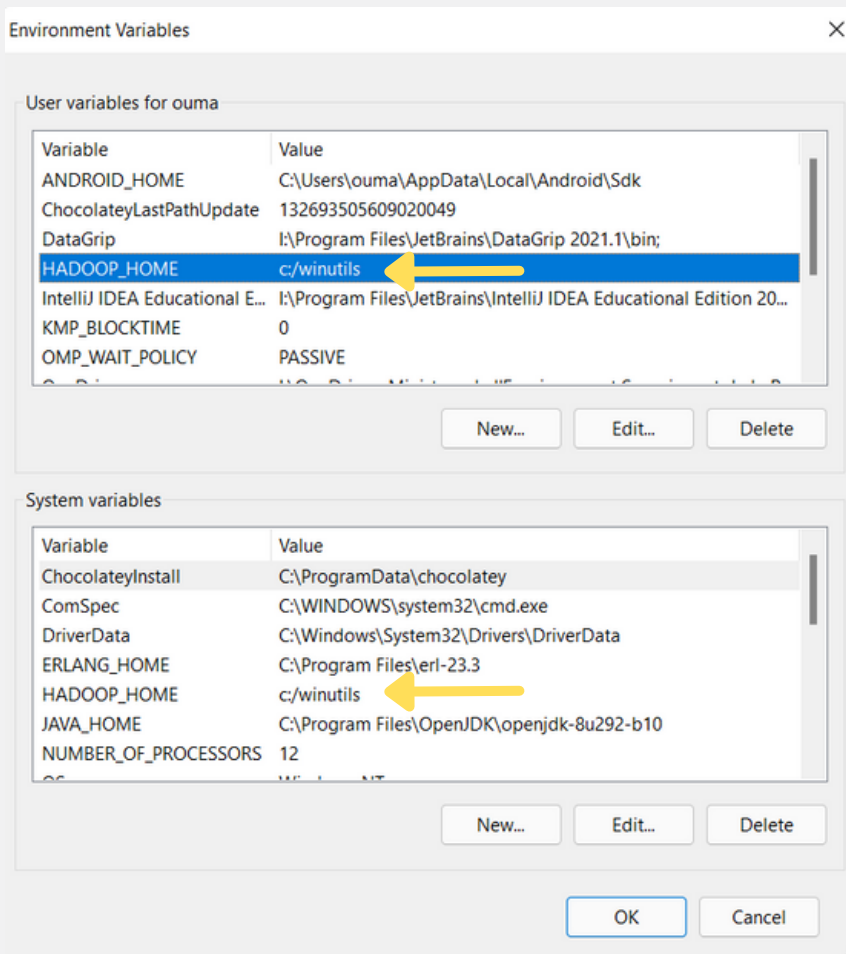
ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tracking UI	Blacklisted Nodes
application_1645104020056_0004	root	purchases count	MAPREDUCE	default	Thu Feb 17 21:06:36 +0100 2022	Thu Feb 17 21:06:58 +0100 2022	FINISHED	SUCCEEDED		History	N/A
application_1645104020056_0002	root	word count	MAPREDUCE	default	Thu Feb 17 20:07:35 +0100 2022	Thu Feb 17 20:09:19 +0100 2022	FINISHED	SUCCEEDED		History	N/A
application_1645104020056_0001	root	word count	MAPREDUCE	default	Thu Feb 17 17:35:59 +0100 2022	Thu Feb 17 17:37:42 +0100 2022	FINISHED	SUCCEEDED		History	N/A

Remarque:

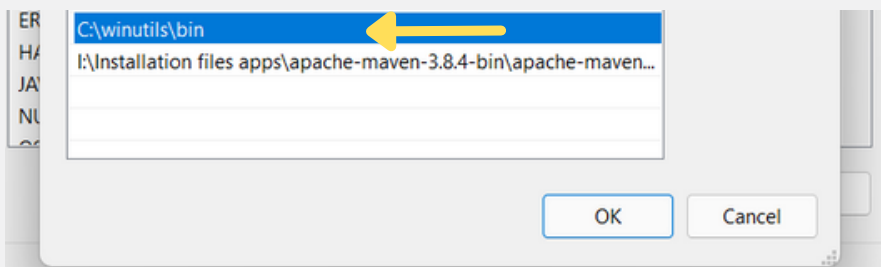
- Il faut ajouter les deux fichiers **hadoop.dll** et **winutils.exe** dans **C:\winutils\bin**
- Pour obtenir les fichiers consulter ce repo:
<https://github.com/cdarlint/winutils/tree/master/hadoop-2.7.2/bin>



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



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



Le code est dans le répo suivant:

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

THE END