

TP1: Installation d'un Cluster Hadoop avec docker

I. Installation Cluster Hadoop

1. Installation docker

```
user@Sara-N:~$ docker version
Client:
Version:           28.4.0
API version:       1.51
Go version:        go1.24.7
Git commit:        d8eb465
Built:             Wed Sep  3 20:56:28 2025
OS/Arch:           linux/amd64
Context:           default

Server: Docker Desktop 4.47.0 (206054)
Engine:
Version:           28.4.0
API version:       1.51 (minimum version 1.24)
Go version:        go1.24.7
Git commit:        249d679
Built:             Wed Sep  3 20:57:37 2025
OS/Arch:           linux/amd64
Experimental:      false
containerd:
Version:           1.7.27
GitCommit:         05044ec0a9a75232cad458027ca83437aae3f4da
runc:
Version:           1.2.5
GitCommit:         v1.2.5-0-g59923ef
docker-init:
Version:           0.19.0
GitCommit:         de40ad0
```

2. Télécharger l'image hadoop-spark-cluste

```
user@Sara-N:~$ docker pull yassern1/hadoop-spark-jupyter:1.0.3
1.0.3: Pulling from yassern1/hadoop-spark-jupyter
4f4fb700ef54: Pull complete
62d02b53c336: Pull complete
7c457f213c76: Pull complete
0c7c8e899c31: Pull complete
88defd485d1b: Pull complete
873ac057ca4e: Pull complete
40091887436b: Pull complete
e44c72237599: Pull complete
93cfcb2f0187: Pull complete
fae462c2169c: Pull complete
b71f1f2e032a: Pull complete
7f61e7eab4ae: Pull complete
3e7b11a95670: Pull complete
b551d0588ff7: Pull complete
70e2e8d7dba3: Pull complete
47fc09f775a8: Pull complete
5c336d7045ea: Pull complete
776a89496dc4: Pull complete
916b439a2e8b: Pull complete
Digest: sha256:511eec146b04568b392ff0a3522c54b47d7c010181334c96bcc9dbbe17ca2fa3
Status: Downloaded newer image for yassern1/hadoop-spark-jupyter:1.0.3
docker.io/yassern1/hadoop-spark-jupyter:1.0.3
```

```
user@Sara-N:~$ docker images
REPOSITORY              TAG          IMAGE ID          CREATED          SIZE
yassern1/hadoop-spark-jupyter  1.0.3       511eec146b04     3 weeks ago     3.67GB
```

3. Création d'un volume de partage

```
user@Sara-N:~$ mkdir -p ~/Documents/hadoop_project
```

```
user@Sara-N:~$ cd Documents
user@Sara-N:~/Documents$ cd hadoop_project
user@Sara-N:~/Documents/hadoop_project$ ls ~/Documents
hadoop_project
```

4. Création des trois conteneurs

Réseau :

```
user@Sara-N:~/Documents/hadoop_project$ docker network create --driver=bridge hadoop
33d9b6dc536bb3ff888ac2156a862372c5b5fb7d4bab97fe037ddf8a14299d90
```

- Conteneur : hadoop-master

```
user@Sara-N:~/Documents/hadoop_project$ docker run -itd -v "$(pwd):/shared_volume" --net=hadoop -p 9870
:9870 -p 8088:8088 -p 7077:7077 -p 19888:19888 --name hadoop-master --hostname hadoop-master yassern1/h
adoop-spark-jupyter:1.0.3
9d70a0fa01360d9d5f7d601f741734b6696cf0c76f9596113788ae179e1b5b24
```

- Conteneur 2 hadoop-slave1

```
user@Sara-N:~/Documents/hadoop_project$ docker run -itd -p 8040:8042 --net=hadoop --name
hadoop-slave1 --hostname hadoop-slave1 yassern1/hadoop-spark-jupyter:1.0.3
43b3b40410dc62f81b8d96f8de3440b82d127862de4e2de7c402959977987df3
```

- Conteneur 3 hadoop-slave2

```
user@Sara-N:~/Documents/hadoop_project$ docker run -itd -p 8041:8042 --net=hadoop --name
hadoop-slave2 --hostname hadoop-slave2 yassern1/hadoop-spark-jupyter:1.0.3
d42bc3356388f24acaee22f6e6fcd33ccd9c1db9ea52c29ea5d2a625ae1e31e9
```

5. Accéder au master

```
user@Sara-N:~/Documents/hadoop_project$ docker exec -it hadoop-master bash
root@hadoop-master:~#
```

6. Démarrer 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: WARNING: HADOOP_NAMENODE_OPTS has been replaced by HDFS_NAMENODE_OPTS. Using value of HADOOP_NAMENODE_OPTS.
Starting datanodes
WARNING: HADOOP_SECURE_DN_LOG_DIR has been replaced by HADOOP_SECURE_LOG_DIR. Using value of HADOOP_SECURE_DN_LOG_DIR.
hadoop-slave2: Warning: Permanently added 'hadoop-slave2,172.18.0.4' (ECDSA) to the list of known hosts.
hadoop-slave1: Warning: Permanently added 'hadoop-slave1,172.18.0.3' (ECDSA) to the list of known hosts.
hadoop-slave1: WARNING: HADOOP_SECURE_DN_LOG_DIR has been replaced by HADOOP_SECURE_LOG_DIR. Using value of HADOOP_SECURE_DN_LOG_DIR.
hadoop-slave1: WARNING: HADOOP_DATANODE_OPTS has been replaced by HDFS_DATANODE_OPTS. Using value of HADOOP_DATANODE_OPTS.
hadoop-slave2: WARNING: HADOOP_SECURE_DN_LOG_DIR has been replaced by HADOOP_SECURE_LOG_DIR. Using value of HADOOP_SECURE_DN_LOG_DIR.
hadoop-slave2: WARNING: HADOOP_DATANODE_OPTS has been replaced by HDFS_DATANODE_OPTS. Using value of HADOOP_DATANODE_OPTS.
Starting secondary namenodes [hadoop-master]
hadoop-master: Warning: Permanently added 'hadoop-master,172.18.0.2' (ECDSA) to the list of known hosts.
hadoop-master: WARNING: HADOOP_SECONDARYNAMENODE_OPTS has been replaced by HDFS_SECONDARYNAMENODE_OPTS. Using value of HADOOP_SECONDARYNAMENODE_OPTS.
Starting resourcemanager
Starting nodemanagers
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.

```

- NameNode web UI: localhost:9870

Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities

Overview 'hadoop-master:9000' (active)

Started:	Mon Oct 06 16:33:07 +0100 2025
Version:	3.2.0, re97acb3bd8f3befd27418996fa5d4b50bf2e17bf
Compiled:	Tue Jan 08 07:08:00 +0100 2019 by sunilg from branch-3.2.0
Cluster ID:	CID-d1633c5c-7365-42a0-a12a-142ec9ca0d73
Block Pool ID:	BP-507101907-127.0.0.1-1757609495422

Summary

Security is off.

Safemode is off.

1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).

Heap Memory used 80.77 MB of 237.5 MB Heap Memory. Max Heap Memory is 836 MB.

Non Heap Memory used 47.37 MB of 48.59 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	1.97 TB
----------------------	---------

- Ressource Manager UI: localhost:8088

hadoop

All Applications
Logged in as: dr.who

Cluster
About
Nodes
Node Labels
Applications
NEW
NEW SAVING
SUBMITTED
ACCEPTED
RUNNING
FINISHED
FAILED
KILLED
Scheduler
Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved
0	0	0	0	0	0 B	4 GB	0 B	0	4	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
2	0	0	0	0	0	0

Scheduler Metrics


Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:2048, vCores:2>	0

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
No data available in table																				

Showing 0 to 0 of 0 entries

- MapReduce JobHistory Server: localhost:19888



JobHistory

Logged in as: dr.who

Application

- About
- Jobs

Tools

Retired Jobs

Show 20 entries
Search:

Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State	Maps Total	Maps Completed	Reduces Total	Reduces Completed	Elapsed Time
No data available in table												

Submit
Start
Finish
Job
Name
User
Queue
State
Maps
Maps Comp
Reduces
Reduces Co
Elapsed

Showing 0 to 0 of 0 entries
First
Previous
Next
Last

7. Manipulations sur HDFS

Créer un dossier input

```
root@hadoop-master:~# hdfs dfs -mkdir /input
```

afficher le contenu de la racine

```
root@hadoop-master:~# hdfs dfs -ls
Found 1 items
drwxr-xr-x - root supergroup 0 2025-10-06 23:09 input
```

Copier le fichier purchases.txt dans le dossier de partage

La création du dossier de partage est sur WSL, donc le fichier purchases.txt doit être transféré depuis windows vers WSL. (/mnt/c/Users/USER/Documents/hadoop_project/)

```
user@Sara-N:~/Documents/hadoop_project$ ls -l /mnt/c/Users/USER/Documents/hadoop_project/purchases.txt
-rwxrwxrwx 1 user user 2549 Oct  6 09:26 /mnt/c/Users/USER/Documents/hadoop_project/purchases.txt
user@Sara-N:~/Documents/hadoop_project$ cp /mnt/c/Users/USER/Documents/hadoop_project/purchases.txt ~/Documents/hadoop_project/
user@Sara-N:~/Documents/hadoop_project$ ls -l ~/Documents/hadoop_project/purchases.txt
-rwxr-xr-x 1 user user 2549 Oct  7 00:36 /home/user/Documents/hadoop_project/purchases.txt
```

Copiez ce fichier sur HDFS

```
root@hadoop-master:~# hdfs dfs -put /shared_volume/purchases.txt .
```

```
root@hadoop-master:~# hdfs dfs -ls -R
drwxr-xr-x - root supergroup 0 2025-10-06 23:09 input
-rw-r--r-- 2 root supergroup 2549 2025-10-06 23:46 purchases.txt
```

Affiche le contenu du fichier

```
root@hadoop-master:~# hdfs dfs -cat purchases.txt
2012-01-01    09:00    San Jose    Men's Clothing    214.05    Amex
2012-01-01    09:00    Fort Worth    Women's Clothing    153.57    Visa
2012-01-01    09:00    San Diego    Music    66.08    Cash
2012-01-01    09:00    Pittsburgh    Pet Supplies    493.51    Discover
2012-01-01    09:00    Omaha    Children's Clothing    235.63    MasterCard
2012-01-01    09:00    Stockton    Men's Clothing    247.18    MasterCard
2012-01-01    09:00    Austin    Cameras    379.6    Visa
2012-01-01    09:00    New York    Consumer Electronics    296.8    Cash
2012-01-01    09:00    Corpus Christi    Toys    25.38    Discover
2012-01-01    09:00    Fort Worth    Toys    213.88    Visa
2012-01-01    09:00    Las Vegas    Video Games    53.26    Visa
2012-01-01    09:00    Newark    Video Games    39.75    Cash
2012-01-01    09:00    Austin    Cameras    469.63    MasterCard
2012-01-01    09:00    Greensboro    DVDs    290.82    MasterCard
2012-01-01    09:00    San Francisco    Music    260.65    Discover
2012-01-01    09:00    Lincoln Garden    136.9    Visa
2012-01-01    09:00    Buffalo    Women's Clothing    483.82    Visa
2012-01-01    09:00    San Jose    Women's Clothing    215.82    Cash
2012-01-01    09:00    Boston    Cameras    418.94    Amex
2012-01-01    09:00    Houston    Baby    309.16    Visa
2012-01-01    09:00    Las Vegas    Books    93.39    Visa
2012-01-01    09:00    Virginia Beach    Children's Clothing    376.11    Amex
2012-01-01    09:01    Riverside    Consumer Electronics    252.88    Cash
2012-01-01    09:01    Tulsa    Baby    205.06    Visa
2012-01-01    09:01    Reno    Crafts    88.25    Visa
2012-01-01    09:01    Chicago    Books    31.08    Cash
2012-01-01    09:01    Fort Wayne    Men's Clothing    370.55    Amex
```

afficher la fin du fichier

```
root@hadoop-master:~# hdfs dfs -tail purchases.txt
ustin    Sporting Goods    327.75    Discover
2012-01-01    09:01    Portland    CDs    108.69    Amex
2012-01-01    09:01    Riverside    Sporting Goods    15.41    Discover
2012-01-01    09:01    Reno    Toys    80.46    Visa
2012-01-01    09:01    Anchorage    Music    298.86    MasterCard
2012-01-01    09:01    Pittsburgh    Sporting Goods    475.26    Amex
2012-01-01    09:01    Spokane    Garden    3.85    Amex
2012-01-01    09:01    Spokane    Computers    287.65    MasterCard
2012-01-01    09:01    Fresno    CDs    466.64    MasterCard
2012-01-01    09:01    Omaha    Baby    255.68    MasterCard
2012-01-01    09:01    Anchorage    DVDs    6.38    Amex
2012-01-01    09:01    Aurora    Consumer Electronics    117.81    MasterCard
2012-01-01    09:01    Philadelphia    DVDs    351.31    Cash
2012-01-01    09:01    Fremont    Baby    222.61    Cash
2012-01-01    09:01    Anchorage    Crafts    22.36    Amex
2012-01-01    09:01    Norfolk    Women's Clothing    180.01    Amex
```

Supprimer ce fichier de HDFS

```
root@hadoop-master:~# hdfs dfs -rm purchases.txt
Deleted purchases.txt
```

```
root@hadoop-master:~# hdfs dfs -ls
Found 1 items
drwxr-xr-x  - root supergroup          0 2025-10-06 23:54 input
```

Remettre à nouveau ce fichier

```
root@hadoop-master:~# hdfs dfs -copyFromLocal /shared_volume/purchases.txt ./input
root@hadoop-master:~# hdfs dfs -ls ./input
Found 1 items
-rw-r--r--  2 root supergroup      2549 2025-10-06 23:54 input/purchases.txt
```

Vérifiez son propriétaire, son groupe et ses droits

```
root@hadoop-master:~# hdfs dfs -chmod go+w ./input/purchases.txt
root@hadoop-master:~# hdfs dfs -chmod go-r ./input/purchases.txt
```

(avant chngement des droits)

```
-rw-r--r--  2 root supergroup      2549 2025-10-06 23:54 input/purchases.txt
```

(après chngement des droits)

```
-rw--w--w-  2 root supergroup      2549 2025-10-06 23:54 input/purchases.txt
```

transférer le fichier de HDFS vers la machine en changeant le nom

```
root@hadoop-master:~# hdfs dfs -get ./input/purchases.txt /shared_volume/achat
root@hadoop-master:~# ls -l /shared_volume
total 8
-rw-r--r--  1 root root 2549 Oct  7 00:06 achat.txt
-rwxr-xr-x  1 1000 1000 2549 Oct  6 23:36 purchases.txt
```

8. télécharger un fichier sur hdf

Créer un dossier input dans le HDFS

```
root@hadoop-master:~# hdfs dfs -mkdir web_input
```

Copier le fichier alice.txt vers le dossier partager hadoop_project

```
user@Sara-N:~$ cp /mnt/c/Users/USER/Documents/hadoop_project/alice.txt ~/Documents/hadoop_project/
user@Sara-N:~$ ls -l ~/Documents/hadoop_project/alice.txt
-rwxr-xr-x  1 user user 150886 Oct  7 15:42 /home/user/Documents/hadoop_project/alice.txt
```

Copier le fichier depuis la machine local vers le dossier input du hdfs

```
root@hadoop-master:~# hdfs dfs -put -f /shared_volume/alice.txt web_input
```

Vérifier l'opération

```
root@hadoop-master:~# hdfs dfs -ls web_input
Found 1 items
-rw-r--r--    2 root supergroup    150886 2025-10-07 14:46 web_input/alice.txt
```

Sortir de bash de hadoop-master

```
root@hadoop-master:~# exit
exit
user@Sara-N:~$
```

Arrêter les trois conteneurs

```
user@Sara-N:~$ docker stop hadoop-master hadoop-slave1 hadoop-slave2
hadoop-master
hadoop-slave1
hadoop-slave2
```