TP 2 : Manipulation d'Apache Hadoop en utilisant des containers Docker et Exécution de programmes MapReduce.

Binome:

-Zahra KASMOUTI

-Imane TOUIBA

1- Installation de Docker sur Ubuntu :

1) Ajout de la clé GPG officielle de Docker :

```
zahra@zahra:/$ curl -fsSL https://download.docker.com/linux/ubuntu/gpg | sudo ap
t-key add -
Warning: apt-key is deprecated. Manage keyring files in trusted.gpg.d instead (s
ee apt-key(8)).
OK
zahra@zahra:/$
```

2) Ajout du dépôt Docker aux sources APT

```
zahra@zahra:/$ sudo add-apt-repository "deb [arch=amd64] https://download.docker
.com/linux/ubuntu $(lsb_release -cs) stable"
Repository: 'deb [arch=amd64] https://download.docker.com/linux/ubuntu oracular
stable'
Description:
Archive for codename: oracular components: stable
More info: https://download.docker.com/linux/ubuntu
Adding repository.
Press [ENTER] to continue or Ctrl-c to cancel.
Adding deb entry to /etc/apt/sources.list.d/archive_uri-https_download_docker_co
m_linux_ubuntu-oracular.list
Adding disabled deb-src entry to /etc/apt/sources.list.d/archive_uri-https_downl
oad_docker_com_linux_ubuntu-oracular.list
Réception de :1 https://download.docker.com/linux/ubuntu oracular InRelease [32,
9 kB]
Réception de :2 https://download.docker.com/linux/ubuntu oracular/stable amd64 P
ackages [5 794 B]
38,6 ko réceptionnés en 5s (7 772 o/s)
Lecture des listes de paquets... Fait
W: https://download.docker.com/linux/ubuntu/dists/oracular/InRelease: Key is sto
red in legacy trusted.gpg keyring (/etc/apt/trusted.gpg), see the DEPRECATION se
ction in apt-key(8) for details.
```

3) Affichage du statut de docker :

```
zahra@zahra:/$ sudo systemctl status docker
● docker.service - Docker Application Container Engine
     Loaded: loaded (/usr/lib/systemd/system/docker.service; enabled; preset: e>
     Active: active (running) since Fri 2024-12-27 09:27:30 UTC; 25s ago
 Invocation: ee3715fc403840ee91a1b96165056f91
TriggeredBy: ● docker.socket
       Docs: https://docs.docker.com
   Main PID: 5515 (dockerd)
      Tasks: 10
     Memory: 34.3M (peak: 35M)
        CPU: 210ms
     CGroup: /system.slice/docker.service
             └─5515 /usr/bin/dockerd -H fd:// --containerd=/run/containerd/cont>

déc. 27 09:27:29 zahra systemd[1]: Starting docker.service - Docker Application>
déc. 27 09:27:29 zahra dockerd[5515]: time="2024-12-27T09:27:29.625206214Z" lev>
déc. 27 09:27:29 zahra dockerd[5515]: time="2024-12-27T09:27:29.625823958Z" lev>
déc. 27 09:27:29 zahra dockerd[5515]: time="2024-12-27T09:27:29.625899750Z" lev>
déc. 27 09:27:29 zahra dockerd[5515]: time="2024-12-27T09:27:29.690091583Z" lev>
déc. 27 09:27:29 zahra dockerd[5515]: time="2024-12-27T09:27:29.976401356Z" lev>
déc. 27 09:27:29 zahra dockerd[5515]: time="2024-12-27T09:27:29.992401804Z" lev>
déc. 27 09:27:29 zahra dockerd[5515]: time="2024-12-27T09:27:29.992505945Z" lev>
```

4) Téléchargement de la version actuelle de Docker Compose :

```
zahra@zahra:/$ sudo curl -L "https://github.com/docker/compose/releases/download
/1.29.2/docker-compose-$(uname -s)-$(uname -m)" -o /usr/local/bin/docker-compose
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
  0     0    0     0    0     0      0      0 --:--:--  0:00:02 --:--:--     0
100 12.1M  100 12.1M    0     0   901k      0  0:00:13  0:00:13 --:--:-- 2412k
zahra@zahra:/$
```

5) Vérification de l'installation :

```
zahra@zahra:/$ docker-compose --version
docker-compose version 1.29.2, build 5becea4c
zahra@zahra:/$
```

6) Tester le bon fonctionnement de Docker :

```
zahra@zahra:/$ sudo docker run hello-world
Messagerie Thunderbird e 'hello-world:latest' locally
latest: Pulling from library/hello-world
c1ec31eb5944: Pull complete
Digest: sha256:5b3cc85e16e3058003c13b7821318369dad01dac3dbb877aac3c28182255c7
Status: Downloaded newer image for hello-world:latest

Hello from Docker!
This message shows that your installation appears to be working correctly.

To generate this message, Docker took the following steps:
 1. The Docker client contacted the Docker daemon.
 2. The Docker daemon pulled the "hello-world" image from the Docker Hub.
    (amd64)
 3. The Docker daemon created a new container from that image which runs the
    executable that produces the output you are currently reading.
 4. The Docker daemon streamed that output to the Docker client, which sent i
    to your terminal.

To try something more ambitious, you can run an Ubuntu container with:
 $ docker run -it ubuntu bash

Share images, automate workflows, and more with a free Docker ID:
```

7) Utilisation de Docker sans sudo :

```
zahra@zahra:/$ sudo chmod 666 /var/run/docker.sock
zahra@zahra:/$ sudo usermod -aG docker $USER
zahra@zahra:/$ sudo systemctl restart docker
zahra@zahra:/$
```

2- Manipulation du HDFS (Commandes de base) :

Accès à l'interface web du NameNode :

localhost:9870/dfshealth.html#tab-overview

Configured Remote Capacity:	0 B
DFS Used:	24 KB (0%)
Non DFS Used:	15.42 GB
DFS Remaining:	11.49 GB (40.48%)
Block Pool Used:	24 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	1 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion (including replicas)	0
Block Deletion Start Time	Fri Dec 27 11:12:45 +0000 2024
Last Checkpoint Time	Fri Dec 27 11:12:44 +0000 2024
Enabled Erasure Coding Policies	RS-6-3-1024k

2- Manipulation du HDFS (Commandes de base) :

2.1- Accès au NameNode, création d'un répertoire dans HDFS "ml-100k" et affichage du contenu de ce répertoire :

```
imane@imane-VirtualBox:~/Downloads/hadoop-main$ docker exec -it namenode /bin/bash
root@0506ef3932bf:/# hdfs dfs -mkdir /ml-100k
root@0506ef3932bf:/# hdfs dfs -ls /
Found 2 items
drwxr-xr-x   - root supergroup          0 2025-01-03 09:54 /ml-100k
drwxr-xr-x   - root supergroup          0 2025-01-03 09:45 /rmstate
root@0506ef3932bf:/#
```

2.2-Copie du fichier poeme.txt depuis le système local vers HDFS :

2.2.1- Copie du poeme.txt depuis l'hôte (hadoop-main) vers le conteneur namenode

```
imane@imane-VirtualBox:~/Downloads/hadoop-main$ docker cp poeme.txt namenode:/
Successfully copied 3.58kB to namenode:/
imane@imane-VirtualBox:~/Downloads/hadoop-main$
```

2.2.2- Accès à nouveau au conteneur Namenode et affichage de son contenu

```
imane@imane-VirtualBox:~/Downloads/hadoop-main$ docker exec -it namenode /bin/bash
root@0506ef3932bf:/# ls
KEYS  boot  entrypoint.sh  hadoop       home  lib32  libx32  mnt  poeme.txt  root  run.sh  srv  tmp  var
bin   dev   etc            hadoop-data  lib   lib64  media   opt  proc       run   sbin    sys  usr
```

2.2.3- Copie du fichier poeme.txt du local du namenode vers le HDFS et affichage du contenu du HDFS :

```
root@0506ef3932bf:/# hdfs dfs -put poeme.txt /
root@0506ef3932bf:/# hdfs dfs -ls /
Found 3 items
drwxr-xr-x   - root supergroup          0 2025-01-03 09:54 /ml-100k
-rw-r--r--   3 root supergroup       1669 2025-01-03 10:02 /poeme.txt
drwxr-xr-x   - root supergroup          0 2025-01-03 09:45 /rmstate
root@0506ef3932bf:/#
```

2.3- Affichage des 20 premières lignes du fichier u.data après avoir copier ce fichier du répertoire local au HDFS :

```
root@0506ef3932bf:/# hdfs dfs -cat /ml-100k/u.data | head -n 20
196     242     3       881250949
186     302     3       891717742
22      377     1       878887116
244     51      2       880606923
166     346     1       886397596
298     474     4       884182806
115     265     2       881171488
253     465     5       891628467
305     451     3       886324817
6       86      3       883603013
62      257     2       879372434
286     1014    5       879781125
200     222     5       876042340
210     40      3       891035994
224     29      3       888104457
303     785     3       879485318
122     387     5       879270459
194     274     2       879539794
291     1042    4       874834944
234     1184    2       892079237
cat: Unable to write to output stream.
root@0506ef3932bf:/#
```

2.4- Affichage des dernières lignes du fichier u.data:

```
root@0506ef3932bf:/# hdfs dfs -tail /ml-100k/u.data
91363685
823     134     5       878438232
130     93      5       874953665
130     121     5       876250746
537     778     3       886031106
655     913     4       891817521
889     2       3       880182460
865     1009    5       880144368
851     979     3       875730244
833     474     5       875122675
394     380     4       881132876
193     690     4       889123221
```

2.5- Renommage du fichier poeme.txt par test.txt et affichage du contenu du répertoire HDFS pour vérifier le renommage du fichier

```
root@0506ef3932bf:/# hdfs dfs -mv /poeme.txt /test.txt
root@0506ef3932bf:/# hdfs dfs -ls /
Found 3 items
drwxr-xr-x   - root supergroup          0 2025-01-03 10:07 /ml-100k
drwxr-xr-x   - root supergroup          0 2025-01-03 09:45 /rmstate
-rw-r--r--   3 root supergroup       1669 2025-01-03 10:02 /test.txt
```

2.6- Téléchargement du fichier test.txt vers le système local

```
root@0506ef3932bf:/# hdfs dfs -get /test.txt /
root@0506ef3932bf:/# ls
KEYS            hadoop       libx32     proc     sys
bin             hadoop-data  media      root     test.txt
boot            home         ml-100k    run      tmp
dev             lib          mnt        run.sh   usr
entrypoint.sh   lib32        opt        sbin     var
etc             lib64        poeme.txt  srv
```

2.7- Suppression le fichier test.txt du HDFS et affichage du contenu de HDFS :

```
root@0506ef3932bf:/# hdfs dfs -rm /test.txt
Deleted /test.txt
root@0506ef3932bf:/# hdfs dfs -ls /
Found 2 items
drwxr-xr-x   - root supergroup          0 2025-01-03 10:07 /ml-100k
drwxr-xr-x   - root supergroup          0 2025-01-03 09:45 /rmstate
```

2.8- Affichage des permissions des répertoires du HDFS:

```
root@0506ef3932bf:/# hdfs dfs -ls /ml-100k/
Found 1 items
-rw-r--r--   3 root supergroup    1979173 2025-01-03 10:07 /ml-100k/u.data
root@0506ef3932bf:/#
```

2.9- Changement des permissions du fichier u.data

```
root@0506ef3932bf:/# hdfs dfs -chmod 777 /ml-100k/u.data
root@0506ef3932bf:/# hdfs dfs -ls /ml-100k/u.data
-rwxrwxrwx   3 root supergroup    1979173 2025-01-03 10:07 /ml-100k/u.data
root@0506ef3932bf:/# hdfs dfs -chmod 755 /ml-100k/u.data
root@0506ef3932bf:/# hdfs dfs -ls /ml-100k/u.data
-rwxr-xr-x   3 root supergroup    1979173 2025-01-03 10:07 /ml-100k/u.data
root@0506ef3932bf:/#
```

2.10- Affichage de la structure du répertoire HDFS

```
root@0506ef3932bf:/# hdfs dfs -du -h /ml-100k/
1.9 M  5.7 M  /ml-100k/u.data
```

2.11- Suppression du répertoire “/ml-100k” et tout son contenu

```
root@0506ef3932bf:/# hdfs dfs -rm -r /ml-100k
Deleted /ml-100k
root@0506ef3932bf:/# hdfs dfs -ls /
Found 1 items
drwxr-xr-x   - root supergroup          0 2025-01-03 09:45 /rmstate
```

2.12- Affichage d’un rapport sur l'utilisation de l'espace disque dans HDFS

```
root@0506ef3932bf:/# hdfs dfsadmin -report
Configured Capacity: 31824109568 (29.64 GB)
Present Capacity: 17876123756 (16.65 GB)
DFS Remaining: 17876041728 (16.65 GB)
DFS Used: 82028 (80.11 KB)
DFS Used%: 0.00%
Replicated Blocks:
        Under replicated blocks: 5
        Blocks with corrupt replicas: 0
        Missing blocks: 0
        Missing blocks (with replication factor 1): 0
        Low redundancy blocks with highest priority to recover: 5
        Pending deletion blocks: 0
Erasure Coded Block Groups:
        Low redundancy block groups: 0
        Block groups with corrupt internal blocks: 0
        Missing block groups: 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0

-------------------------------------------------
Live datanodes (1):

Name: 172.18.0.6:9866 (datanode.hadoop-main_default)
Hostname: 9036b8e185f8
Decommission Status : Normal
Configured Capacity: 31824109568 (29.64 GB)
DFS Used: 82028 (80.11 KB)
Non DFS Used: 12305444756 (11.46 GB)
DFS Remaining: 17876041728 (16.65 GB)
DFS Used%: 0.00%
DFS Remaining%: 56.17%
Configured Cache Capacity: 0 (0 B)
```

3- Exécution du problème MapReduce « Word Count » en Java avec un cluster constitué d'un Namenode et d'un Datanode:

3.1- Dépôt de répertoire code dans le container namenode :

```
imane@imane-VirtualBox:~/Downloads/hadoop-main$ docker cp codeWordCount namenode:/

Successfully copied 9.22kB to namenode:/
```

3.2- Copie du fichier "poeme.txt" dans HDFS :

```
root@0506ef3932bf:/# hdfs dfs -put poeme.txt /
root@0506ef3932bf:/# hdfs dfs -ls /
Found 2 items
-rw-r--r--   3 root supergroup       1669 2025-01-03 10:31 /poeme.txt
drwxr-xr-x   - root supergroup          0 2025-01-03 09:45 /rmstate
root@0506ef3932bf:/#
```

3.3- Après compilation des classes java , génération et exécution du .jar : Affichage du contenu du fichier

/results/part-r-00000 contenant les résultats de la tâche MapReduce.

```
root@0506ef3932bf:/codeWordCount# hadoop fs -cat /results/part-r-00000
a        6 occurences.
adoraient        1 occurences.
ailes    1 occurences.
aima     1 occurences.
amour    1 occurences.
au       11 occurences.
bas      1 occurences.
belle    1 occurences.
bles     1 occurences.
bras     1 occurences.
bretagne         1 occurences.
brula    1 occurences.
celle    1 occurences.
celui    20 occurences.
cette    1 occurences.
chancelle        1 occurences.
chapelle         1 occurences.
ciel     10 occurences.
citadelle        1 occurences.
clarte   1 occurences.
coeur    2 occurences.
combat   1 occurences.
comment 1 occurences.
commun   1 occurences.
coule    2 occurences.
```

3.4- Arrêt et suppression de tous les conteneurs existants

```
imane@imane-VirtualBox:~/Downloads/hadoop-main$ docker-compose down
Stopping nodemanager     ... done
Stopping resourcemanager ... done
Stopping namenode        ... done
Stopping datanode        ... done
Stopping historyserver   ... done
Removing nodemanager     ... done
Removing resourcemanager ... done
Removing namenode        ... done
Removing datanode        ... done
Removing historyserver   ... done
Removing network hadoop-main_default
imane@imane-VirtualBox:~/Downloads/hadoop-main$
```

3.5- Affichage des volumes utilisés par Docker

```
imane@imane-VirtualBox:~/Downloads/hadoop-main$ docker volume ls
DRIVER     VOLUME NAME
local      hadoop-main_hadoop_datanode
local      hadoop-main_hadoop_historyserver
local      hadoop-main_hadoop_namenode
```

3.6- Inspection du volume “hadoop-main_hadoop_namenode”

```
imane@imane-VirtualBox:~/Downloads/hadoop-main$ docker volume inspect hadoop-main_hadoop_namenode

[
    {
        "CreatedAt": "2025-01-03T10:37:33+01:00",
        "Driver": "local",
        "Labels": {
            "com.docker.compose.project": "hadoop-main",
            "com.docker.compose.version": "1.29.2",
            "com.docker.compose.volume": "hadoop_namenode"
        },
        "Mountpoint": "/var/lib/docker/volumes/hadoop-main_hadoop_namenode/_data",
        "Name": "hadoop-main_hadoop_namenode",
        "Options": null,
        "Scope": "local"
    }
]
imane@imane-VirtualBox:~/Downloads/hadoop-main$
```

4- Configuration d’un cluster multi-datanodes (un namenode et de deux datanodes) et Exécution du problème MapReduce “WordCount” en java”:

Étape 1: Lancer les conteneurs (namenode et datanodes) :

Étape 2: Préparation des fichiers nécessaires :

Étape 3:Vérification de l'état des nœuds et des services Hadoop:

4.1- voir l'état des nœuds HDFS

```
root@79c2d3facca6:/# hdfs dfsadmin -report
Configured Capacity: 63648219136 (59.28 GB)
Present Capacity: 35782323748 (33.32 GB)
DFS Remaining: 35782270976 (33.32 GB)
DFS Used: 52772 (51.54 KB)
DFS Used%: 0.00%
Replicated Blocks:
        Under replicated blocks: 6
        Blocks with corrupt replicas: 0
        Missing blocks: 0
        Missing blocks (with replication factor 1): 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0
Erasure Coded Block Groups:
        Low redundancy block groups: 0
        Block groups with corrupt internal blocks: 0
        Missing block groups: 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0

-------------------------------------------------
Live datanodes (2):

Name: 172.18.0.7:9866 (datanode2.hadoop-main_default)
Hostname: 8c5617aec8dd
Decommission Status : Normal
Configured Capacity: 31824109568 (29.64 GB)
DFS Used: 26386 (25.77 KB)
Non DFS Used: 12290406638 (11.45 GB)
```

4.2- Soumettre le .jar et l'exécuter

```
root@79c2d3facca6:/codeWordCount# hadoop jar wcount.jar org.hadoop.wordcount.WCount /poeme.txt /result2
2025-01-03 12:04:17,263 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/172.18.0.2:8032
2025-01-03 12:04:17,736 INFO client.AHSProxy: Connecting to Application History server at historyserver/172.18.0.4:10200
2025-01-03 12:04:18,343 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/
root/.staging/job_1735904840163_0002
2025-01-03 12:04:19,140 INFO input.FileInputFormat: Total input files to process : 1
2025-01-03 12:04:19,552 INFO mapreduce.JobSubmitter: number of splits:1
2025-01-03 12:04:20,065 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1735904840163_0002
2025-01-03 12:04:20,066 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-01-03 12:04:20,509 INFO conf.Configuration: resource-types.xml not found
2025-01-03 12:04:20,533 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-01-03 12:04:20,987 INFO impl.YarnClientImpl: Submitted application application_1735904840163_0002
2025-01-03 12:04:21,150 INFO mapreduce.Job: The url to track the job: http://resourcemanager:8088/proxy/application_1735
904840163_0002/
2025-01-03 12:04:21,164 INFO mapreduce.Job: Running job: job_1735904840163_0002
2025-01-03 12:04:33,797 INFO mapreduce.Job: Job job_1735904840163_0002 running in uber mode : false
2025-01-03 12:04:33,807 INFO mapreduce.Job:  map 0% reduce 0%
2025-01-03 12:04:43,284 INFO mapreduce.Job:  map 100% reduce 0%
2025-01-03 12:04:51,530 INFO mapreduce.Job:  map 100% reduce 100%
2025-01-03 12:04:52,564 INFO mapreduce.Job: Job job_1735904840163_0002 completed successfully
2025-01-03 12:04:52,778 INFO mapreduce.Job: Counters: 54
```

4.3- Affichage du contenu du fichier /result2/part-r-00000 contenant les résultats de la tâche MapReduce

```
root@79c2d3facca6:/codeWordCount#  hdfs dfs -ls /
Found 6 items
drwxrwxrwt   - root root                0 2025-01-03 12:04 /app-logs
-rw-r--r--   3 root supergroup       1669 2025-01-03 11:48 /poeme.txt
drwxr-xr-x   - root supergroup          0 2025-01-03 12:04 /result2
drwxr-xr-x   - root supergroup          0 2025-01-03 11:59 /results2
drwxr-xr-x   - root supergroup          0 2025-01-03 11:47 /rmstate
drwx------   - root supergroup          0 2025-01-03 11:58 /tmp
root@79c2d3facca6:/codeWordCount# hadoop fs -cat /result2/part-r-00000
a       6 occurences.
adoraient       1 occurences.
ailes   1 occurences.
aima    1 occurences.
amour   1 occurences.
au      11 occurences.
bas     1 occurences.
belle   1 occurences.
bles    1 occurences.
bras    1 occurences.
bretagne        1 occurences.
brula   1 occurences.
celle   1 occurences.
celui   20 occurences.
cette   1 occurences.
chancelle       1 occurences.
chapelle        1 occurences.
ciel    10 occurences.
citadelle       1 occurences.
clarte  1 occurences.
coeur   2 occurences.
```