Bilan de l’installation

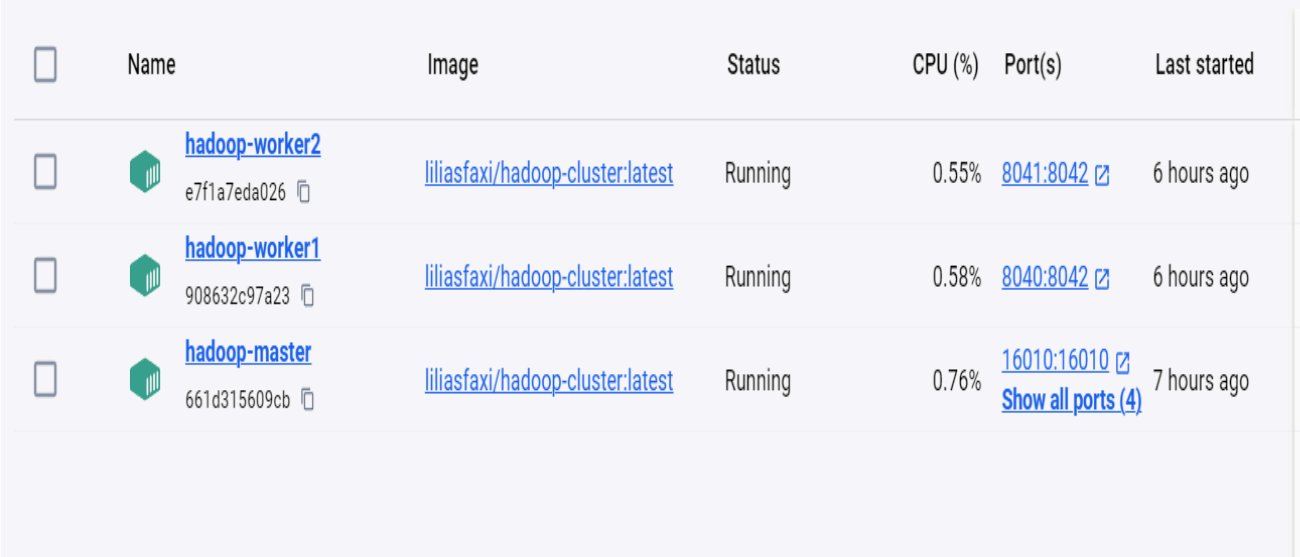
Introduction : présentation de l’objectif :

Mise en place d'un environnement Hadoop avec Docker Desktop pour l'analyse de données volumineuses

Ce guide détaille les étapes nécessaires pour installer Docker Desktop, exécuter une image Hadoop et gérer des conteneurs Docker afin d'effectuer l'analyse de données volumineuses. Le fichier `purchases.txt` servira de base de données pour illustrer le processus. Ceci a comme but de calculer le nombre de ventes par magasin et la vente totale par magasin.

2)Installation de Docker Desktop:

- installation Docker Desktop en téléchargeant le logiciel depuis le site officiel de Docker.
-Téléchargement de l'image de Hadoop qui porte le nom de liliasfaxi/spark-hadoop:hv-2.7.2



Name	Image	Status	CPU (%)	Port(s)	Last started
hadoop-worker2 e7f1a7eda026	liliasfaxi/hadoop-cluster:latest	Running	0.55%	8041:8042	6 hours ago
hadoop-worker1 908632c97a23	liliasfaxi/hadoop-cluster:latest	Running	0.58%	8040:8042	6 hours ago
hadoop-master 661d315609cb	liliasfaxi/hadoop-cluster:latest	Running	0.76%	16010:16010 Show all ports (4)	7 hours ago

3)Téléchargement de l’image de hadoop et la création +lancement des 3 contenaires :

-J'ai utilisé la commande suivante pour télécharger l'image :

```
docker pull liliasfaxi/hadoop-cluster:latest
```

```
C:\Users\Amen Khlifi> docker pull liliasfaxi/spark-hadoop:hv-2.
hv-2.7.2: Pulling from liliasfaxi/spark-hadoop
1be7f2b886e8: Pull complete
6fbc4a21b806: Pull complete
c71a6f8e1378: Pull complete
4be3072e5a37: Pull complete
06c6d2f59700: Pull complete
b8606274051a: Pull complete
8176485c06ce: Pull complete
f3a132dac987: Pull complete
a3c7183d2677: Pull complete
d010f061a722: Pull complete
d81c164d96f9: Pull complete
d8d441090d24: Pull complete
7c12d721deef: Pull complete
091d1ad175e0: Pull complete
793a639c13bb: Pull complete
040b0d6351fa: Pull complete
262437b95da7: Pull complete
Digest: sha256:56f4243e1b22684301e611df6e724605846f4ddbaf8d8884
Status: Downloaded newer image for liliasfaxi/spark-hadoop:hv-2
docker.io/liliasfaxi/spark-hadoop:hv-2.7.2
```

-J'ai crée un réseau qui permettra de relier les trois contenaires avec la commande suivante :

docker network create --driver=bridge hadoop

-création et lancement des trois contenaires :hadoop-master , hadoop-worke1 , hadoop-worker2

```
C:\Users\Amen Khlifi> docker ps
CONTAINER ID   IMAGE                          COMMAND                  CREATED         STATUS
                                                NAMES
e7f1a7eda026   liliasfaxi/hadoop-cluster:latest   "sh -c 'service ssh …"   14 seconds ago   Up 13 sec
                                                     hadoop-worker2
908632c97a23   liliasfaxi/hadoop-cluster:latest   "sh -c 'service ssh …"   54 seconds ago   Up 54 sec
                                                     hadoop-worker1
661d315609cb   liliasfaxi/hadoop-cluster:latest   "sh -c 'service ssh …"   44 minutes ago   Up 44 min
0.0.0.0:9870->9870/tcp, 0.0.0.0:16010->16010/tcp   hadoop-master
```

Me déplacer dans le contenaire master pour commencer à l'utiliser.

docker exec -it hadoop-master bash

4)exécution de Hadoop et yarn avec la manipulation de fichiers dans HDFS :

lancement du hadoop et yarn:

./start-hadoop.sh

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


Starting namenodes on [hadoop-master]
hadoop-master: Warning: Permanently added 'hadoop-master' (ED25519) to th
ist of known hosts.
hadoop-master: WARNING: HADOOP_NAMENODE_OPTS has been replaced by HDFS_NA
ODE_OPTS. Using value of HADOOP_NAMENODE_OPTS.
Starting datanodes
WARNING: HADOOP_SECURE_DN_LOG_DIR has been replaced by HADOOP_SECURE_LOG_
. Using value of HADOOP_SECURE_DN_LOG_DIR.
hadoop-worker1: Warning: Permanently added 'hadoop-worker1' (ED25519) to
 list of known hosts.
hadoop-worker2: Warning: Permanently added 'hadoop-worker2' (ED25519) to
 list of known hosts.
hadoop-worker1: WARNING: HADOOP_SECURE_DN_LOG_DIR has been replaced by HA
P_SECURE_LOG_DIR. Using value of HADOOP_SECURE_DN_LOG_DIR.
hadoop-worker1: WARNING: HADOOP_DATANODE_OPTS has been replaced by HDFS_D
NODE_OPTS. Using value of HADOOP_DATANODE_OPTS.
hadoop-worker2: WARNING: HADOOP_SECURE_DN_LOG_DIR has been replaced by HA
P_SECURE_LOG_DIR. Using value of HADOOP_SECURE_DN_LOG_DIR.
hadoop-worker2: WARNING: HADOOP_DATANODE_OPTS has been replaced by HDFS_D
NODE_OPTS. Using value of HADOOP_DATANODE_OPTS.
Starting secondary namenodes [hadoop-master]
hadoop-master: Warning: Permanently added 'hadoop-master' (ED25519) to th
ist of known hosts.
hadoop-master: WARNING: HADOOP_SECONDARYNAMENODE_OPTS has been replaced b
DFS_SECONDARYNAMENODE_OPTS. Using value of HADOOP_SECONDARYNAMENODE_OPTS.


Starting resourcemanager
Starting nodemanagers
hadoop-worker1: Warning: Permanently added 'hadoop-worker1' (ED25519) to
 list of known hosts.
hadoop-worker2: Warning: Permanently added 'hadoop-worker2' (ED25519) to
 list of known hosts.
```

-Créer un répertoire dans HDFS, appelé *input:*

Hadoop fs -mkdir -p input

Chargement du fichier purchases dans le répertoire input :

Hadoop fs -put purchases.txt input

Afficher le contenu de input :

Hadoop fs -ls input

```
C:\Users\Amen Khlifi>docker exec -it hadoop-master bash
root@hadoop-master:~# hadoop fs -mkdir -p input
root@hadoop-master:~# ls
hdfs  purchases.txt  run-wordcount.sh  start-hadoop.sh  start-kafka-zookee
root@hadoop-master:~# hadoop fs -ls
Found 1 items
drwxr-xr-x   - root supergroup          0 2024-04-21 16:37 input
root@hadoop-master:~# hadoop fs -ls input
Found 1 items
drwxr-xr-x   - root supergroup          0 2024-04-21 16:37 input/purchases
root@hadoop-master:~#
```

Pour afficher les dernières lignes du fichier purchases.txt situé dans le répertoire input, j'ai exécuté la commande suivante :

Hadoop fs -tail 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
```

5)Passons à mapreduce :

MapReduce est un modèle de programmation et un framework logiciel utilisé pour traiter des ensembles de données volumineuses de manière parallèle et distribuée. Il s'agit d'un concept fondamental dans le traitement du Big Data, en particulier lorsque l'on travaille avec Apache Hadoop.

Phase Map :

Preparation du fichier mapper :

mapper.py × reducer.py /usr/local/hadoop/share/hadoop/tools/lib

C: > Users > Amen Khlifi > OneDrive > Bureau > mapper.py > ...

```python
#!/usr/bin/python
# Format of each line is:
# date\ttime\tstore name\titem description\tcost\tmethod of
#
# We want elements 2 (store name) and 4 (cost)
# We need to write them out to standard output, separated by
import sys
for line in sys.stdin:
    data = line.strip().split("\t")
    if len(data) == 6:
        date, time, store, item, cost, payment = data
        print("{0}\t{1}".format(store, cost))
```

Preparation du fichier reducer :

mapper.py | reducer.py | /usr/local/hadoop/share/hadoop/tools/lib Untitled-

C: > Users > Amen Khlifi > OneDrive > Bureau > reducer.py > ...

```python
#!/usr/bin/python
# Format of each line is:
# date\ttime\tstore name\titem description\tcost\tmethod of payme
#
# We want elements 2 (store name) and 4 (cost)
# We need to write them out to standard output, separated by a ta
import sys
salesTotal = 0
oldKey = None
# Loop around the data
# It will be in the format key\tval
# Where key is the store name, val is the sale amount
#
# All the sales for a particular store will be presented,
# then the key will change and we'll be dealing with the next sto
for line in sys.stdin:
    data_mapped = line.strip().split("\t")
    if len(data_mapped) != 2:
        # Something has gone wrong. Skip this line.
        continue
    thisKey, thisSale = data_mapped
    if oldKey and oldKey != thisKey:
        print(oldKey, "\t", salesTotal)
        oldKey = thisKey
        salesTotal = 0
    oldKey = thisKey
    salesTotal += float(thisSale)
if oldKey != None:
    print(oldKey, "\t", salesTotal)
```

Charger le mapper et reducer dans hadoop master :

```
C:\Users\Amen Khlifi>docker cp "C:\Users\Amen Khlifi\OneDr
Successfully copied 2.05kB to hadoop-master:/root/mapper.p

C:\Users\Amen Khlifi>docker cp "C:\Users\Amen Khlifi\OneDr
Successfully copied 2.56kB to hadoop-master:/root/reducer.
```

Lancement de hadoop streaming jar :

```
root@hadoop-master:/usr/bin# cd ~
root@hadoop-master:~# hadoop jar /usr/local/hadoop/share/hadoop/tools/lib
-input input/purchases.txt \
-output /output2 \
-mapper "python3 mapper.py" \
-reducer "python3 reducer.py" \
-file mapper.py \
-file reducer.py
```

Apàres l’execution de mapper et reducer on peut visionner le resultat de cette technique dans le fichier output2 :

```
root@hadoop-master:~# ls
hdfs  mapper.py  purchases.txt  reducer.py  run-wordcount.sh  start-hadoo
root@hadoop-master:~# hdfs dfs -ls
Found 1 items
drwxr-xr-x   - root supergroup          0 2024-04-21 16:37 input
root@hadoop-master:~# hdfs dfs -ls /output2
Found 2 items
-rw-r--r--   2 root supergroup          0 2024-04-23 08:43 /output2/_SUCC
-rw-r--r--   2 root supergroup       3100 2024-04-23 08:43 /output2/part-
root@hadoop-master:~# hdfs dfs -tail /output2/part-00000
```

Resultat du tail :

```
root@hadoop-master:~# hdfs dfs -tail /output2/part-00000
maha      10026642.340000028
Orlando           10074922.52000003
Philadelphia      10190080.259999994
Phoenix           10079076.699999955
Pittsburgh        10090124.82000001
Plano     10046103.60999996
Portland          10007635.770000052
Raleigh           10061442.539999973
Reno      10079955.16000006
Richmond          9992941.590000007
Riverside         10006695.41999998
Rochester         10067606.920000048
Sacramento        10123468.180000057
Saint Paul        10057233.569999998
San Antonio       10014441.700000012
San Bernardino    9965152.040000001
San Diego         9966038.390000042
San Francisco     9995570.540000083
San Jose          9936721.410000023
Santa Ana         10050309.929999964
Scottsdale        10037929.850000067
Seattle           9936267.369999954
Spokane           10083362.980000008
St. Louis         10002105.139999984
St. Petersburg    9986495.54
Stockton          10006412.639999997
Tampa     10106428.550000126
Toledo    10020768.879999934
```

6)Erreur que j’ai rencontré :

Dans le lancement de hadoop jar :

```
root@hadoop-master:/usr/local/hadoop/share/hadoop/tools/lib# cd ~
root@hadoop-master:~# hadoop jar /usr/local/hadoop/share/hadoop/tools/lib
    -input /input/purchases.txt \
    -output /output/result \
    -mapper "python mapper.py" \
    -reducer "python reducer.py" \
    -file mapper.py \
    -file reducer.py \
2024-04-23 08:34:57,224 WARN streaming.StreamJob: -file option is depreca
packageJobJar: [mapper.py, reducer.py, /tmp/hadoop-unjar7700575080199932
mpDir=null
2024-04-23 08:34:57,998 INFO client.DefaultNoHARMFailoverProxyProvider: C
72.18.0.4:8032
2024-04-23 08:34:58,192 INFO client.DefaultNoHARMFailoverProxyProvider: C
72.18.0.4:8032
2024-04-23 08:34:58,392 INFO mapreduce.JobResourceUploader: Disabling Era
root/.staging/job_1713859969569_0001
2024-04-23 08:34:59,368 INFO mapreduce.JobSubmitter: Cleaning up the stag
/job_1713859969569_0001
2024-04-23 08:34:59,388 ERROR streaming.StreamJob: Error Launching job :
:9000/input/purchases.txt
Streaming Command Failed!
root@hadoop-master:~# hadoop jar /usr/local/hadoop/share/hadoop/tools/lib
   -output /output/result     -mapper "python mapper.py"     -reducer "py
reducer.py \
>
```

Au premieu lieu j’ai ecrit la commande de cette manière qui est fausse alors j’ai commencé à chercher le path de hadoop jar , le nom de saisie de python comme le montre les images ci-dessous :

Chercher python :

```
root@hadoop-master:~# cd /usr/bin/python
bash: cd: /usr/bin/python: No such file or directory
root@hadoop-master:~# cd /usr/bin/
root@hadoop-master:/usr/bin# ls
 X11                                gdk-pixbuf-pixdata       openssl
'['                                 gdk-pixbuf-thumbnailer   orbd
 addpart                            getconf                  pack200
 appletviewer                       getent                   pager
 appres                             getopt                   partx
 apt                                gpasswd                  passwd
 apt-cache                          gpgv                     paste
 apt-cdrom                          grep                     pathchk
 apt-config                         groups                   pdb3
 apt-get                            gtk-update-icon-cache    pdb3.10
 apt-key                            gunzip                   perl
 apt-mark                           gzexe                    perl5.34.
 arch                               gzip                     pgrep
 awk                                hardlink                 pidof
 b2sum                              head                     pidwait
 base32                             helpztags                pinky
 base64                             hostid                   pkill
 basename                           hostname                 pldd
 basenc                             hostnamectl              pmap
 bash                               hsdb                     policytoo
 bashbug                            i386                     pr
 bootctl                            iconv                    printenv
 busctl                             id                       printf
```

```
root@hadoop-master:/usr/bin# cd python3
bash: cd: python3: Not a directory
root@hadoop-master:/usr/bin# cd python3.
bash: cd: python3.: No such file or directory
root@hadoop-master:/usr/bin# cd python3.10
bash: cd: python3.10: Not a directory
root@hadoop-master:/usr/bin# ls -l
total 41432
lrwxrwxrwx 1 root root         1 Mar 25  2022  X11 -> .
-rwxr-xr-x 1 root root     51632 Feb  7  2022 '['
-rwxr-xr-x 1 root root     14712 Feb 21  2022  addpart
lrwxrwxrwx 1 root root        30 Jan  1 14:24  appletviewer -> /etc/alte
-rwxr-xr-x 1 root root     14648 Mar 25  2022  appres
-rwxr-xr-x 1 root root     18824 Sep 28  2022  apt
-rwxr-xr-x 1 root root     84448 Sep 28  2022  apt-cache
-rwxr-xr-x 1 root root     27104 Sep 28  2022  apt-cdrom
-rwxr-xr-x 1 root root     27024 Sep 28  2022  apt-config
-rwxr-xr-x 1 root root     51680 Sep 28  2022  apt-get
-rwxr-xr-x 1 root root     28173 Sep 28  2022  apt-key
-rwxr-xr-x 1 root root     51680 Sep 28  2022  apt-mark
-rwxr-xr-x 1 root root     31232 Feb  7  2022  arch
lrwxrwxrwx 1 root root        21 Nov  1  2022  awk -> /etc/alternatives/
-rwxr-xr-x 1 root root     51720 Feb  7  2022  b2sum
-rwxr-xr-x 1 root root     35328 Feb  7  2022  base32
-rwxr-xr-x 1 root root     35328 Feb  7  2022  base64
-rwxr-xr-x 1 root root     35328 Feb  7  2022  basename
-rwxr-xr-x 1 root root     47616 Feb  7  2022  basenc
-rwxr-xr-x 1 root root   1396520 Jan  6  2022  bash
-rwxr-xr-x 1 root root      6818 Jan  6  2022  bashbug
```

Chercher hadoop streaming pour le lancer :

```
root@hadoop-master:/usr/local/hadoop/bin# ls -l
total 1736
-rwxr-xr-x 1 1000 1000 802984 Jun 18  2023 container-executor
-rwxr-xr-x 1 1000 1000   9036 Jun 18  2023 hadoop
-rwxr-xr-x 1 1000 1000  11265 Jun 18  2023 hadoop.cmd
-rwxr-xr-x 1 1000 1000  11274 Jun 18  2023 hdfs
-rwxr-xr-x 1 1000 1000   8081 Jun 18  2023 hdfs.cmd
-rwxr-xr-x 1 1000 1000   6349 Jun 18  2023 mapred
-rwxr-xr-x 1 1000 1000   6311 Jun 18  2023 mapred.cmd
-rwxr-xr-x 1 1000 1000  33448 Jun 18  2023 oom-listener
-rwxr-xr-x 1 1000 1000 837112 Jun 18  2023 test-container-executor
-rwxr-xr-x 1 1000 1000  12439 Jun 18  2023 yarn
-rwxr-xr-x 1 1000 1000  12840 Jun 18  2023 yarn.cmd
root@hadoop-master:/usr/local/hadoop/bin# cd ~
root@hadoop-master:~# locate hadoop-streaming.jar
bash: locate: command not found
root@hadoop-master:~# sudo updatedb
bash: sudo: command not found
root@hadoop-master:~# /usr/local/hadoop/share/tools/lib/hadoop-streaming-
bash: /usr/local/hadoop/share/tools/lib/hadoop-streaming-x.y.z.jar: No su
root@hadoop-master:~# cd /usr/local/hadoop/share/tools/lib/hadoop-streami
bash: cd: /usr/local/hadoop/share/tools/lib/hadoop-streaming-3.3.6.jar: N
root@hadoop-master:~# cd usr/local/hadoop/share/tools/lib/hadoop-streamin
bash: cd: usr/local/hadoop/share/tools/lib/hadoop-streaming-3.3.6.jar: No
root@hadoop-master:~# /usr/local/hadoop/share/tools/lib/hadoop-streaming-
bash: /usr/local/hadoop/share/tools/lib/hadoop-streaming-3.3.6.jar: No su
```

```
root@hadoop-master:~# cd /usr
root@hadoop-master:/usr# ls
bin  games  include  lib  lib32  lib64  libexec  libx32  local  sbin  sha
root@hadoop-master:/usr# cd loczl
bash: cd: loczl: No such file or directory
root@hadoop-master:/usr# cd local
root@hadoop-master:/usr/local# ls
bin  etc  games  hadoop  hbase  include  kafka  lib  man  sbin  share  sp
root@hadoop-master:/usr/local# cd hadoop
root@hadoop-master:/usr/local/hadoop# ls
LICENSE-binary  NOTICE-binary  README.txt  etc      lib      licenses-bi
LICENSE.txt     NOTICE.txt     bin         include  libexec  logs
root@hadoop-master:/usr/local/hadoop# cd share
root@hadoop-master:/usr/local/hadoop/share# ls
doc  hadoop
root@hadoop-master:/usr/local/hadoop/share# cd hadoop
root@hadoop-master:/usr/local/hadoop/share/hadoop# ls*
bash: ls*: command not found
root@hadoop-master:/usr/local/hadoop/share/hadoop# ls
client  common  hdfs  mapreduce  tools  yarn
root@hadoop-master:/usr/local/hadoop/share/hadoop# cd tools
root@hadoop-master:/usr/local/hadoop/share/hadoop/tools# ls
dynamometer  lib  resourceestimator  sls  sources
root@hadoop-master:/usr/local/hadoop/share/hadoop/tools# cd lib
root@hadoop-master:/usr/local/hadoop/share/hadoop/tools/lib# ls
aliyun-java-sdk-core-4.5.10.jar      hadoop-datajoin-3.3.6.jar
aliyun-java-sdk-kms-2.11.0.jar       hadoop-distcp-3.3.6.jar
```

Et je me suis rendu compte que la correcte commande s'écrit de cette façon :

hadoop jar /usr/local/share/Hadoop/tools/lib/hadoop-streaming-3.3.6 jar \

-input input/purchases.txt\

-output/output2\

-mapper "python3 mapper.py"\

-reducer "python3 reducer.py"\

-file mapper.py

-file reducer.py