

Big Data

Compte Rendu TP1

I- Docker

- Définition :

Docker est un outil qui permet de créer, gérer et exécuter des applications de manière isolée dans des conteneurs. Un conteneur Docker est un environnement autonome qui contient une application et toutes ses dépendances, ce qui le rend portable, léger et facile à déployer sur différentes plateformes. En simplifiant, Docker permet d'emballer une application avec tout ce dont elle a besoin pour fonctionner, de manière à ce qu'elle puisse être exécutée de manière cohérente et prévisible, quel que soit l'endroit où vous la déployez

- Image Docker :

Une image Docker est un modèle ou un gabarit qui contient tous les éléments nécessaires à l'exécution d'une application dans un conteneur Docker. Plus précisément, une image Docker est une capture figée d'un système de fichiers, y compris le code de l'application, les bibliothèques, les dépendances et les configurations, ainsi que des informations sur la manière dont l'application doit être exécutée.

- Container Docker :

Un conteneur Docker est une instance exécutable d'une image Docker. Il s'agit d'un environnement léger et isolé qui contient tout ce dont une application a besoin pour fonctionner, y compris le code de l'application, les bibliothèques, les dépendances et les configurations.

- Docker Compose :

Docker Compose est un outil de la suite Docker qui permet de définir, configurer et exécuter des applications multi-conteneurs de manière simple et efficace. Il est principalement utilisé pour orchestrer plusieurs conteneurs Docker ensemble, en décrivant la structure de l'application et ses dépendances dans un fichier YAML appelé "docker-compose.yml".

- La commande qui crée les images Docker à partir d'un fichier Docker Compose est :
> **docker-compose build**

II-Notre hadoop cluster :

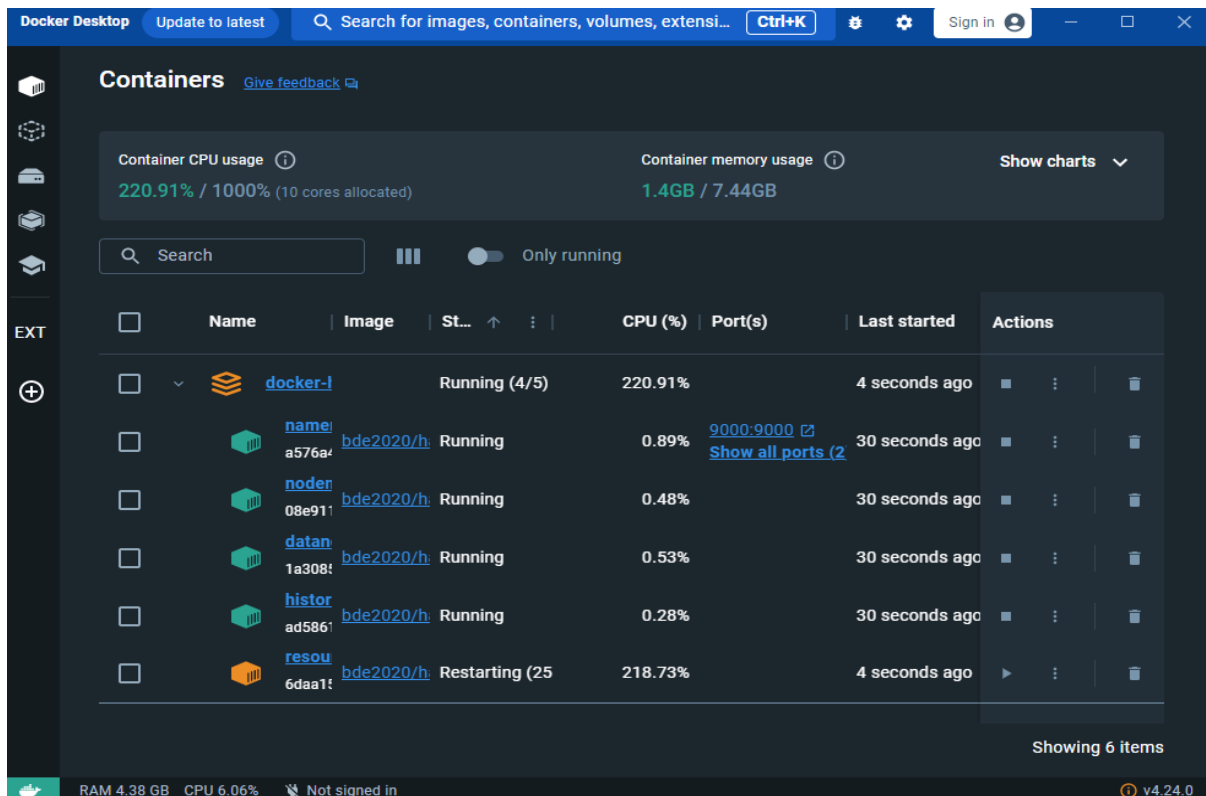
```

1 version: "3"
2
3 services:
4   namenode:
5     image: bde2020/hadoop-namenode:2.0.0-hadoop3.2.1-java8
6     container_name: namenode
7     restart: always
8     ports:
9       - 9870:9870
10      - 9000:9000
11     volumes:
12       - hadoop_namenode:/hadoop/dfs/name
13     environment:
14       - CLUSTER_NAME=test
15     env_file:
16       - ./hadoop.env
17
18   datanode:
19     image: bde2020/hadoop-datanode:2.0.0-hadoop3.2.1-java8
20     container_name: datanode
21     restart: always
22     volumes:
23       - hadoop_datanode:/hadoop/dfs/data
24     environment:
25       SERVICE_PRECONDITION: "namenode:9870"
26     env_file:
27       - ./hadoop.env
28
29   resourcemanager:
30     image: bde2020/hadoop-resourcemanager:2.0.0-hadoop3.2.1-java8
31     container_name: resourcemanager
32     restart: always
33     environment:
34       SERVICE_PRECONDITION: "namenode:9000 namenode:9870 datanode:9864"
35     env_file:
36       - ./hadoop.env
37
38   nodemanager1:
39     image: bde2020/hadoop-nodemanager:2.0.0-hadoop3.2.1-java8
40     container_name: nodemanager
41     restart: always
42     environment:
43       SERVICE_PRECONDITION: "namenode:9000 namenode:9870 datanode:9864 resou"
44     env_file:
45       - ./hadoop.env
46
47   historyserver:
48     image: bde2020/hadoop-historyserver:2.0.0-hadoop3.2.1-java8
49     container_name: historyserver
50     restart: always
51     environment:
52       SERVICE_PRECONDITION: "namenode:9000 namenode:9870 datanode:9864 resou"
53     volumes:
54       - hadoop_historyserver:/hadoop/yarn/timeline
55     env_file:
56       - ./hadoop.env
57
58 volumes:
59   hadoop_namenode:
60   hadoop_datanode:
61   hadoop_historyserver:
  
```

Le cluster est composé de 5 container :

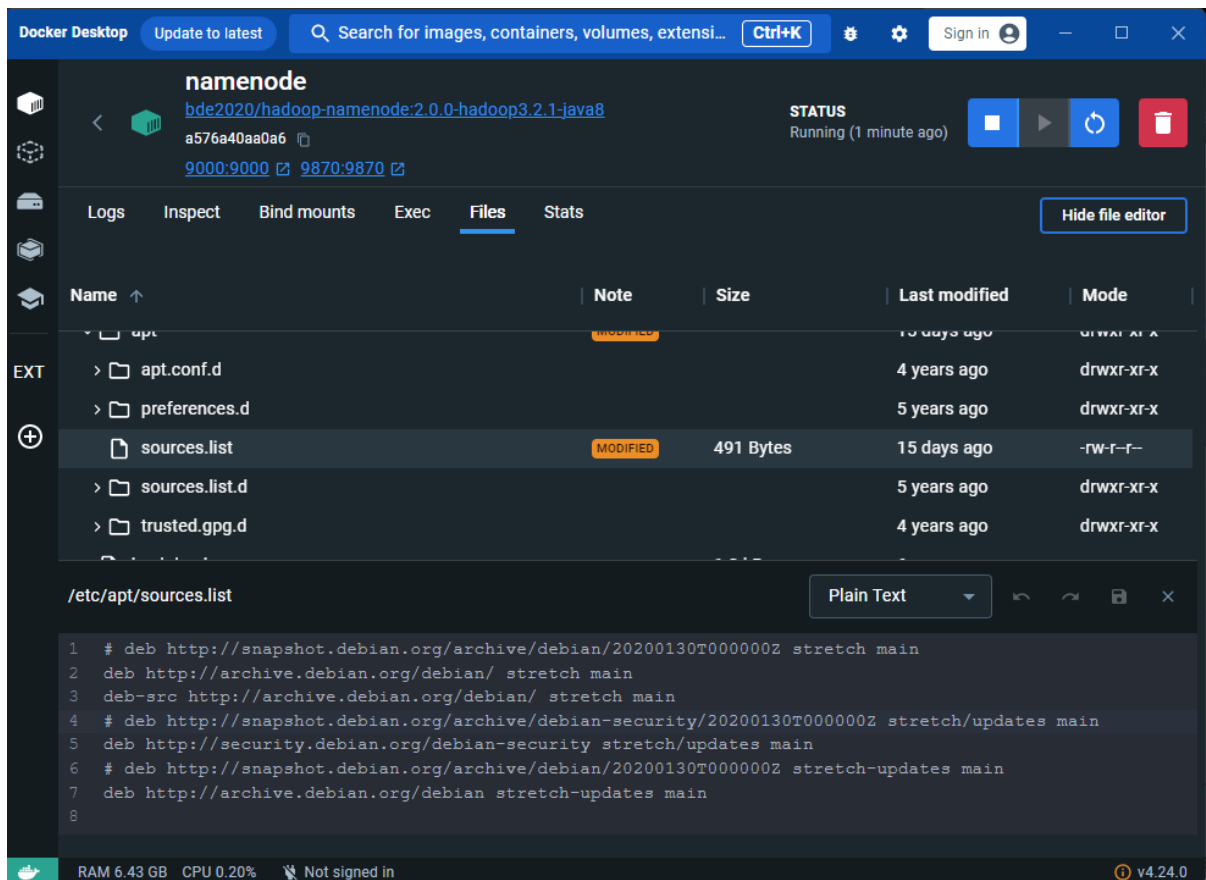
- namenode
- datanode
- resourcemanager
- nodemanager
- historyserver

=> Après l'exécution de docker compose build on a eu les container suivantes :



- Installation de python sur tous les machines :

1- Tout d'abord il faut modifier les lien que le SE debian utilise pour faire ses mise a jours donc on accède sur chaque machine à : **/apt/source.list** et on introduit le contenu suivant :



2- on ouvre accède à chaque machine directement via la fenêtre exec ou dans powershell en exécutant la commande suivante :

> **docker exec -it container_name bash**

3 - mettre à jour la liste de paquet disponible sur chaque machine en exécutant la commande suivante

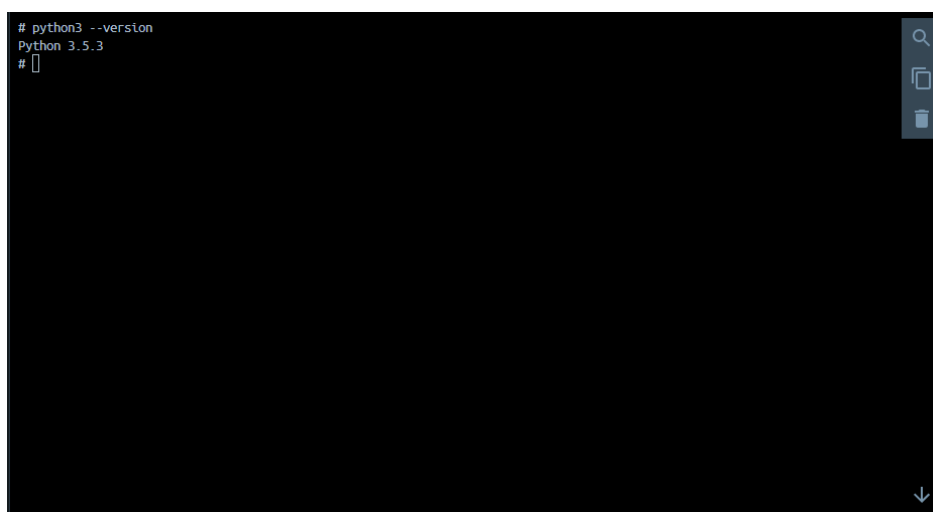
> **apt-get update**

3 - mettre à jour la liste de paquet disponible sur chaque machine en exécutant la commande suivante

> **apt-get install python3**

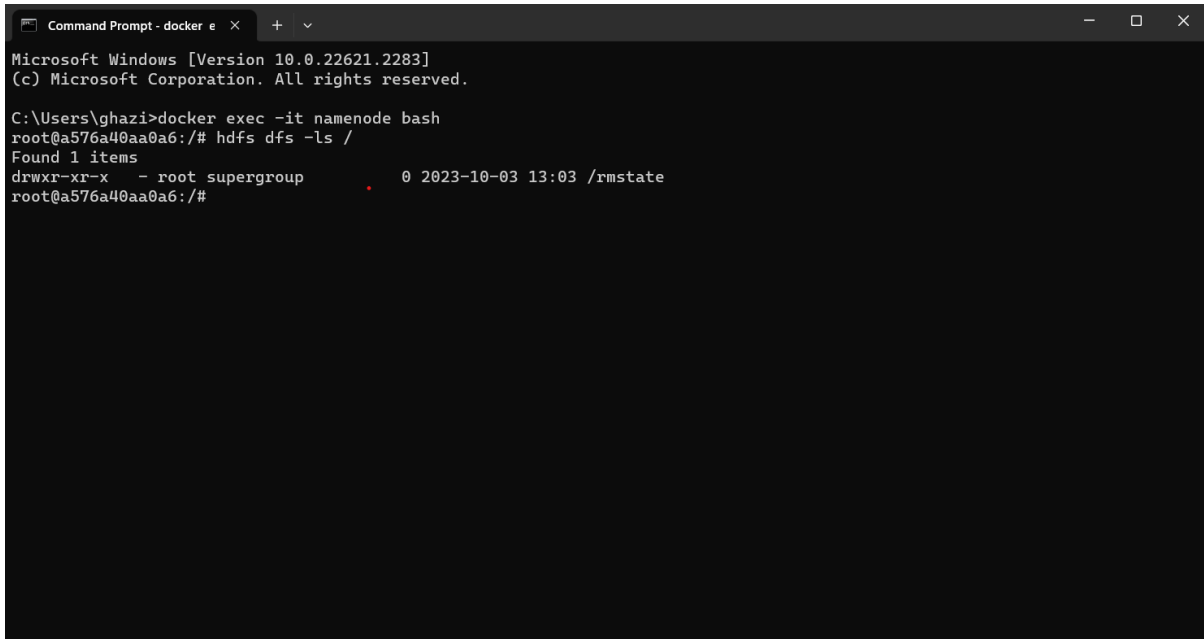
4 - vérification de l'installation de python :

> **python3 --version**



III- Commande HDFS

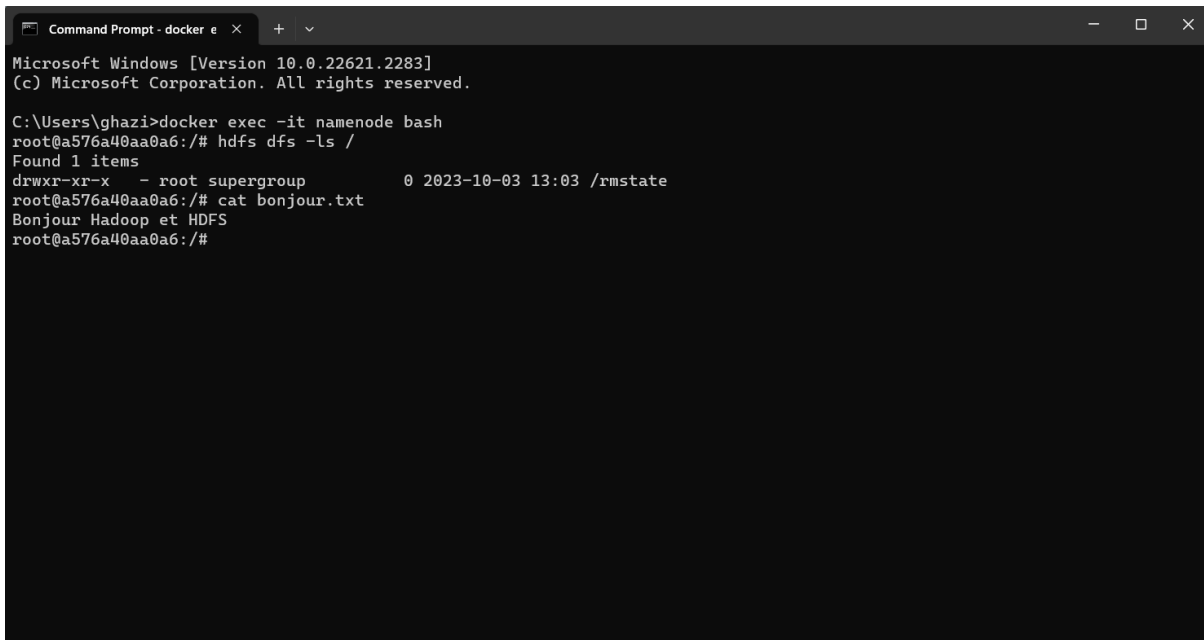
- Accéder à la machine namenode avec docker exec



```
Microsoft Windows [Version 10.0.22621.2283]
(c) Microsoft Corporation. All rights reserved.

C:\Users\ghazi>docker exec -it namenode bash
root@a576a40aa0a6:/# hdfs dfs -ls /
Found 1 items
drwxr-xr-x  - root supergroup          0 2023-10-03 13:03 /rmstate
root@a576a40aa0a6:/#
```

- Création du fichier bonjour.txt et affichage de son contenu



```
Microsoft Windows [Version 10.0.22621.2283]
(c) Microsoft Corporation. All rights reserved.

C:\Users\ghazi>docker exec -it namenode bash
root@a576a40aa0a6:/# hdfs dfs -ls /
Found 1 items
drwxr-xr-x  - root supergroup          0 2023-10-03 13:03 /rmstate
root@a576a40aa0a6:/# cat bonjour.txt
Bonjour Hadoop et HDFS
root@a576a40aa0a6:/#
```

- Création d'une repertoire nommé dossier dans hdfs

```
Command Prompt - docker e X + v
Microsoft Windows [Version 10.0.22621.2283]
(c) Microsoft Corporation. All rights reserved.

C:\Users\ghazi>docker exec -it namenode bash
root@a576a40aa0a6:/# hdfs dfs -ls /
Found 1 items
drwxr-xr-x - root supergroup 0 2023-10-03 13:03 /rmstate
root@a576a40aa0a6:/# cat bonjour.txt
Bonjour Hadoop et HDFS
root@a576a40aa0a6:/# hdfs dfs -ls -R -h /var
ls: '/var': No such file or directory
root@a576a40aa0a6:/# hdfs dfs -mkdir dossier
mkdir: 'hdfs://namenode:9000/user/root': No such file or directory
root@a576a40aa0a6:/# hdfs dfs -ls /
Found 1 items
drwxr-xr-x - root supergroup 0 2023-10-03 13:03 /rmstate
root@a576a40aa0a6:/# hdfs dfs -mkdir /dossier
root@a576a40aa0a6:/# hdfs dfs -ls /
Found 2 items
drwxr-xr-x - root supergroup 0 2023-10-10 14:15 /dossier
drwxr-xr-x - root supergroup 0 2023-10-03 13:03 /rmstate
root@a576a40aa0a6:/#
```

- Ajout de fichier bonjour.txt sous hdfs

```
Command Prompt - docker e X + v
root@a576a40aa0a6:/# cat bonjour.txt
Bonjour Hadoop et HDFS
root@a576a40aa0a6:/# hdfs dfs -put bonjour.txt
put: '.': No such file or directory: 'hdfs://namenode:9000/user/root'
root@a576a40aa0a6:/# hdfs dfs -put bonjour.txt /
2023-10-10 14:17:16,193 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
root@a576a40aa0a6:/# hdfs dfs -ls /
Found 3 items
-rw-r--r-- 3 root supergroup 23 2023-10-10 14:17 /bonjour.txt
drwxr-xr-x - root supergroup 0 2023-10-10 14:15 /dossier
drwxr-xr-x - root supergroup 0 2023-10-03 13:03 /rmstate
root@a576a40aa0a6:/# hdfs dfs -ls -R
ls: '.': No such file or directory
root@a576a40aa0a6:/# hdfs dfs -ls -R /
-rw-r--r-- 3 root supergroup 23 2023-10-10 14:17 /bonjour.txt
drwxr-xr-x - root supergroup 0 2023-10-10 14:15 /dossier
drwxr-xr-x - root supergroup 0 2023-10-03 13:03 /rmstate
drwxr-xr-x - root supergroup 0 2023-10-10 13:58 /rmstate/FSRMStateRoot
drwxr-xr-x - root supergroup 0 2023-10-03 13:03 /rmstate/FSRMStateRoot/AMRMTOKENSecretManagerRoot
-rw-r--r-- 3 root supergroup 23 2023-10-03 13:03 /rmstate/FSRMStateRoot/AMRMTOKENSecretManagerRoot/AMRMTOKENSecretManagerNode
-rw-r--r-- 3 root supergroup 2 2023-10-10 13:58 /rmstate/FSRMStateRoot/EpochNode
drwxr-xr-x - root supergroup 0 2023-10-03 13:03 /rmstate/FSRMStateRoot/RMAppRoot
drwxr-xr-x - root supergroup 0 2023-10-10 13:58 /rmstate/FSRMStateRoot/RMDTSecretManagerRoot
-rw-r--r-- 3 root supergroup 17 2023-10-03 13:03 /rmstate/FSRMStateRoot/RMDTSecretManagerRoot/DelegationKey_1
-rw-r--r-- 3 root supergroup 17 2023-10-03 13:03 /rmstate/FSRMStateRoot/RMDTSecretManagerRoot/DelegationKey_2
-rw-r--r-- 3 root supergroup 17 2023-10-08 09:06 /rmstate/FSRMStateRoot/RMDTSecretManagerRoot/DelegationKey_3
-rw-r--r-- 3 root supergroup 17 2023-10-08 09:06 /rmstate/FSRMStateRoot/RMDTSecretManagerRoot/DelegationKey_4
-rw-r--r-- 3 root supergroup 17 2023-10-08 09:07 /rmstate/FSRMStateRoot/RMDTSecretManagerRoot/DelegationKey_5
```

- Afficher son contenu sous hdfs

```
Command Prompt - docker e X + v
drwxr-xr-x - root supergroup 0 2023-10-03 13:03 /rmstate/FSRMStateRoot/AMRMTOKENSecretManagerRoot
-rw-r--r-- 3 root supergroup 23 2023-10-03 13:03 /rmstate/FSRMStateRoot/AMRMTOKENSecretManagerRoot/AMRMTOKENSe
cretManagerNode
-rw-r--r-- 3 root supergroup 2 2023-10-10 13:58 /rmstate/FSRMStateRoot/EpochNode
drwxr-xr-x - root supergroup 0 2023-10-03 13:03 /rmstate/FSRMStateRoot/RMAppRoot
drwxr-xr-x - root supergroup 0 2023-10-10 13:58 /rmstate/FSRMStateRoot/RMDTSecretManagerRoot
-rw-r--r-- 3 root supergroup 17 2023-10-03 13:03 /rmstate/FSRMStateRoot/RMDTSecretManagerRoot/DelegationKey_1
-rw-r--r-- 3 root supergroup 17 2023-10-03 13:03 /rmstate/FSRMStateRoot/RMDTSecretManagerRoot/DelegationKey_2
-rw-r--r-- 3 root supergroup 17 2023-10-08 09:06 /rmstate/FSRMStateRoot/RMDTSecretManagerRoot/DelegationKey_3
-rw-r--r-- 3 root supergroup 17 2023-10-08 09:06 /rmstate/FSRMStateRoot/RMDTSecretManagerRoot/DelegationKey_4
-rw-r--r-- 3 root supergroup 17 2023-10-08 09:07 /rmstate/FSRMStateRoot/RMDTSecretManagerRoot/DelegationKey_5
-rw-r--r-- 3 root supergroup 17 2023-10-08 09:07 /rmstate/FSRMStateRoot/RMDTSecretManagerRoot/DelegationKey_6
-rw-r--r-- 3 root supergroup 17 2023-10-10 13:58 /rmstate/FSRMStateRoot/RMDTSecretManagerRoot/DelegationKey_7
-rw-r--r-- 3 root supergroup 17 2023-10-10 13:58 /rmstate/FSRMStateRoot/RMDTSecretManagerRoot/DelegationKey_8
-rw-r--r-- 3 root supergroup 4 2023-10-03 13:03 /rmstate/FSRMStateRoot/RMVersionNode
drwxr-xr-x - root supergroup 0 2023-10-03 13:03 /rmstate/FSRMStateRoot/ReservationSystemRoot
root@a576a40aa0a6:/# hdfs dfs -mkdir -p /root
root@a576a40aa0a6:/# hdfs dfs -ls
bash: hdfs: command not found
root@a576a40aa0a6:/# hdfs dfs -ls
ls: '.': No such file or directory
root@a576a40aa0a6:/# hdfs dfs -cat /bonjour.txt | more
2023-10-10 14:21:05,895 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
Bonjour Hadoop et HDFS
root@a576a40aa0a6:/# hdfs dfs -tail /bonjour.txt
2023-10-10 14:22:25,998 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
Bonjour Hadoop et HDFS
root@a576a40aa0a6:/#
```

- Utilisation de commande copyFromLocal pour remettre le fichier sous hdfs, changer ses droit d'accès et le mettre dans la répertoire dossier créé au début de tp

```
Command Prompt - docker e X + v
root@a576a40aa0a6:/# hdfs dfs -rm /bonjour.txt
Deleted /bonjour.txt
root@a576a40aa0a6:/# hdfs dfs -copyFromLocal bonjour.txt
copyFromLocal: '.': No such file or directory: 'hdfs://namenode:9000/user/root'
root@a576a40aa0a6:/# hdfs dfs -copyFromLocal /bonjour.txt
copyFromLocal: '.': No such file or directory: 'hdfs://namenode:9000/user/root'
root@a576a40aa0a6:/# hdfs dfs -copyFromLocal /root/bonjour.txt
copyFromLocal: '.': No such file or directory: 'hdfs://namenode:9000/user/root'
root@a576a40aa0a6:/# hdfs dfs -copyFromLocal bonjour.txt
copyFromLocal: '.': No such file or directory: 'hdfs://namenode:9000/user/root'
root@a576a40aa0a6:/# hdfs dfs -copyFromLocal bonjour.txt /
2023-10-10 14:26:22,297 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
root@a576a40aa0a6:/# hdfs dfs -ls /
Found 4 items
-rw-r--r-- 3 root supergroup 23 2023-10-10 14:26 /bonjour.txt
drwxr-xr-x - root supergroup 0 2023-10-10 14:15 /dossier
drwxr-xr-x - root supergroup 0 2023-10-03 13:03 /rmstate
drwxr-xr-x - root supergroup 0 2023-10-10 14:20 /root
root@a576a40aa0a6:/# hdfs dfs -chmod go+w /bonjour.txt
root@a576a40aa0a6:/# hdfs dfs -ls /
Found 4 items
-rw-rw-rw- 3 root supergroup 23 2023-10-10 14:26 /bonjour.txt
drwxr-xr-x - root supergroup 0 2023-10-10 14:15 /dossier
drwxr-xr-x - root supergroup 0 2023-10-03 13:03 /rmstate
drwxr-xr-x - root supergroup 0 2023-10-10 14:20 /root
root@a576a40aa0a6:/# hdfs dfs -chmod go+r /bonjour.txt
root@a576a40aa0a6:/# hdfs dfs -ls /
Found 4 items
-rw-rw-rw- 3 root supergroup 23 2023-10-10 14:26 /bonjour.txt
drwxr-xr-x - root supergroup 0 2023-10-10 14:15 /dossier
drwxr-xr-x - root supergroup 0 2023-10-03 13:03 /rmstate
drwxr-xr-x - root supergroup 0 2023-10-10 14:20 /root
root@a576a40aa0a6:/# hdfs dfs -mv /bonjour.txt /dossier/bonjour.txt
root@a576a40aa0a6:/#
```

- Copier le fichier a partir de hdfs vers le local et le nommer bien.txt

```
root@a576a40aa0a6:/# hdfs dfs -mv /bonjour.txt /dossier/bonjour.txt
root@a576a40aa0a6:/# hdfs dfs -get /dossier/bonjour.txt /dossier/bien.txt
get: '/dossier/bien.txt': No such file or directory: 'file:///dossier/bien.txt'
root@a576a40aa0a6:/# hdfs dfs -get /dossier/bonjour.txt bien.txt
2023-10-10 14:33:16,140 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
root@a576a40aa0a6:/# ls
KEYS bin boot entrypoint.sh hadoop home lib64 mnt proc run sbin sys usr
bien.txt bonjour.txt dev etc hadoop-data lib media opt root run.sh srv tmp var
root@a576a40aa0a6:/#
```

- Copier le fichier dans le répertoire dossier de hdfs sous le nom de bien.txt

```
root@a576a40aa0a6:/# hdfs dfs -cp /dossier/bonjour.txt /dossier/bien.txt
2023-10-10 14:36:05,663 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2023-10-10 14:36:05,780 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
root@a576a40aa0a6:/# hdfs dfs -ls /dossier
Found 2 items
-rw-r--r-- 3 root supergroup 23 2023-10-10 14:36 /dossier/bien.txt
-rw-rw-rw- 3 root supergroup 23 2023-10-10 14:26 /dossier/bonjour.txt
root@a576a40aa0a6:/#
```

- Affichage de nombre de sous dossier, fichier dans la répertoire dossier et le supprimer de hdfs

```
root@a576a40aa0a6:/# hdfs dfs -count -h /dossier
      1      2      46 /dossier
root@a576a40aa0a6:/# hdfs dfs -rm /dossier/bonjour.txt
Deleted /dossier/bonjour.txt
root@a576a40aa0a6:/# hdfs dfs -rm -r /dossier
Deleted /dossier
root@a576a40aa0a6:/# hdfs dfs -ls
ls: `.`: No such file or directory
root@a576a40aa0a6:/# hdfs dfs -ls /
Found 2 items
drwxr-xr-x - root supergroup      0 2023-10-03 13:03 /rmstate
drwxr-xr-x - root supergroup      0 2023-10-10 14:20 /root
root@a576a40aa0a6:/#
```

VI- Création d'une arborescence et téléchargement de fichier

- Création de l'arborescence

```
Command Prompt - docker e x + v
root@a576a40aa0a6:/# hdfs dfs -mkdir /Tps
root@a576a40aa0a6:/# hdfs dfs -mkdir /Tps/data
root@a576a40aa0a6:/# hdfs dfs -mkdir /Tps/code
root@a576a40aa0a6:/# hdfs dfs -ls /Tps
Found 2 items
drwxr-xr-x - root supergroup      0 2023-10-10 14:46 /Tps/code
drwxr-xr-x - root supergroup      0 2023-10-10 14:46 /Tps/data
root@a576a40aa0a6:/#
```

- Mettre le fichier sous namenode après son téléchargement en local

```
Command Prompt - docker e x Windows PowerShell x + v
PS C:\Users\ghazi> docker cp "C:\Users\ghazi\Downloads\purchases (1).txt.gz" namenode:/
Successfully copied 38.5MB to namenode:/
PS C:\Users\ghazi>
```

- Télécharger le fichier directement à partir de namenode en utilisant curl et affichage de resultat

```
Command Prompt - docker e X Windows PowerShell X + v
root@a576a40aa0a6:/# hdfs dfs -ls /Tps/data
Found 2 items
-rw-r--r-- 3 root supergroup 3369248 2023-10-15 12:17 /Tps/data/pg135.txt
-rw-r--r-- 3 root supergroup 38454568 2023-10-15 12:10 /Tps/data/purchases.txt.gz
root@a576a40aa0a6:/#
```