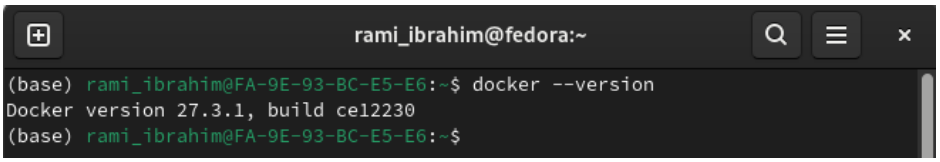
> Rami Ibrahim ING 2 TD 2
> Compte rendu TP1 BIG DATA

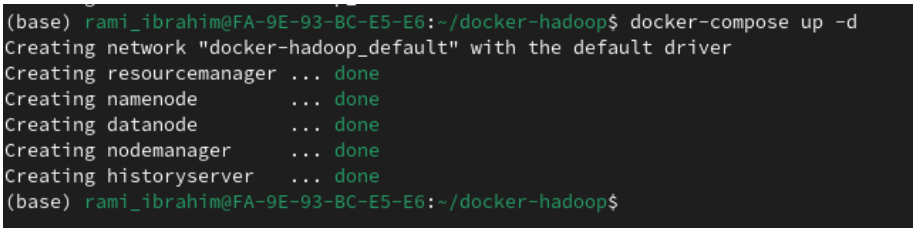
In order to work with hadoop , we need to set it up with Docker or natively.
Since i am comfortable with Docker, i can use prebuilt Hadoop Docker images or set up my own containerized Hadoop cluster.

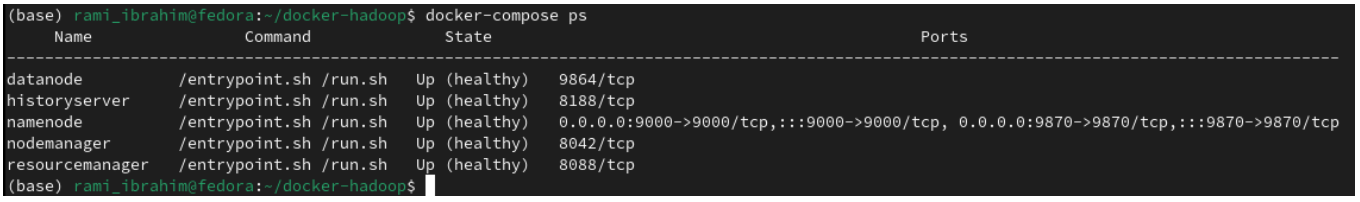
Step 1: Install Docker



Step 2: Clone the git repo

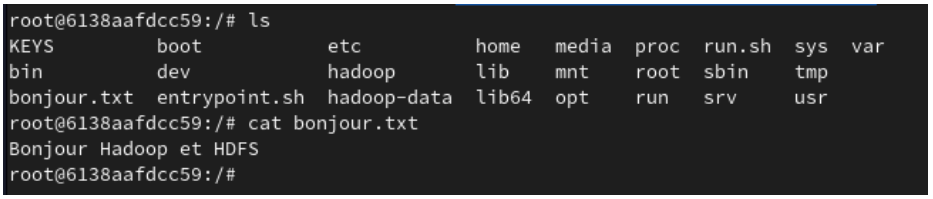
**Step 3: Run the cluster

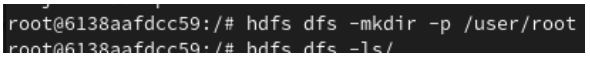


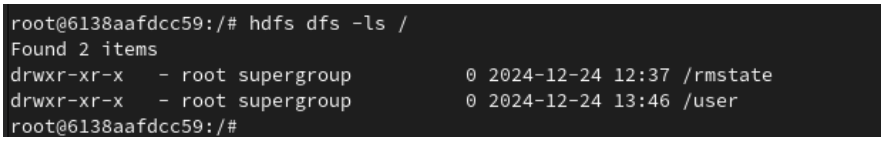


Connect to container namenode

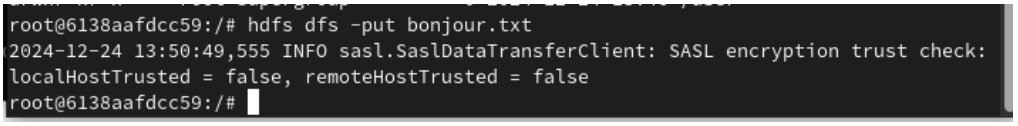
Create a file `bonjour.txt`







Now let's copy `bonjour.txt` to the HDFS



Let's verify

```
root@6138aafdcc59:/# hdfs dfs -ls /user/root
Found 1 items
-rw-r--r--   3 root supergroup         23 2024-12-24 13:50 /user/root/bonjour.txt
root@6138aafdcc59:/#
```

Let's also check the content of the copied file

```
root@6138aafdcc59:/# hdfs dfs -cat bonjour.txt
2024-12-24 13:52:29,260 INFO sasl.SaslDataTransferClient: SASL encryption trust check:
localHostTrusted = false, remoteHostTrusted = false
Bonjour Hadoop et HDFS
root@6138aafdcc59:/#
```

Let's create the structure of our repository with a parent folder `TPs`

```
rami_ibrahim@fedora:~/docker-hadoop — docker exec -it namenode b...
rami_ibrahim@fedora:~/docker-hadoo...   rami_ibrahim@fedora:~/docker-hadoo...
root@6138aafdcc59:/# hdfs dfs -mkdir -p /TPs/data
root@6138aafdcc59:/# hdfs dfs -ls /
Found 3 items
drwxr-xr-x   - root supergroup          0 2024-12-24 13:55 /TPs
drwxr-xr-x   - root supergroup          0 2024-12-24 12:37 /rmstate
drwxr-xr-x   - root supergroup          0 2024-12-24 13:46 /user
root@6138aafdcc59:/#
```

After creating the structure , shift back to the native and download `purchases.txt`

```
(base) rami_ibrahim@fedora:~/docker-hadoop$ curl -o purchases.txt https://www.kaggle.com/datasets/dsfelix/purchasestxt
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  9532    0  9532    0     0  19527      0 --:--:-- --:--:-- --:--:-- 19532
(base) rami_ibrahim@fedora:~/docker-hadoop$
```

And then copy the file into the container `namenode`

```
(base) rami_ibrahim@fedora:~/docker-hadoop$ docker cp purchases.txt namenode:/root
Successfully copied 11.3kB to namenode:/root
(base) rami_ibrahim@fedora:~/docker-hadoop$
```

Then back to the HDFS and copy `purchases.txt`

```
root@6138aafdcc59:/# hdfs dfs -put /root/purchases.txt /TPs/data
2024-12-24 14:06:07,376 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
root@6138aafdcc59:/#
```

Again same process for page 4300 and page 135

```
(base) rami_ibrahim@fedora:~/docker-hadoop$ docker exec namenode curl -o pg4300.txt https://www.gutenberg.org/cache/epub/4300/pg4300.txt
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 1549k  100 1549k    0     0   358k      0  0:00:04  0:00:04 --:--:--  358k
(base) rami_ibrahim@fedora:~/docker-hadoop$
```

```
(base) rami_ibrahim@fedora:~/docker-hadoop$ docker exec namenode hdfs dfs -put pg4300.txt /TPs/data
2024-12-24 14:19:20,895 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
```

```
(base) rami_ibrahim@fedora:~/docker-hadoop$ curl -o pg135.txt https://www.gutenberg.org/cache/epub/135/pg135.txt
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 3290k  100 3290k    0     0   445k      0  0:00:07  0:00:07 --:--:--  538k
(base) rami_ibrahim@fedora:~/docker-hadoop$
```

At the end we find the 3 files copied into the HDFS

```
root@6138aafdcc59:~# hdfs dfs -ls /TPs/data
Found 3 items
-rw-r--r--   3 root supergroup    3369250 2024-12-24 14:34 /TPs/data/pg135.txt
-rw-r--r--   3 root supergroup    1586382 2024-12-24 14:19 /TPs/data/pg4300.txt
-rw-r--r--   3 root supergroup       9532 2024-12-24 14:06 /TPs/data/purchases.txt
root@6138aafdcc59:~#
```

State of the cluster