

Rapport du 04/04/2024

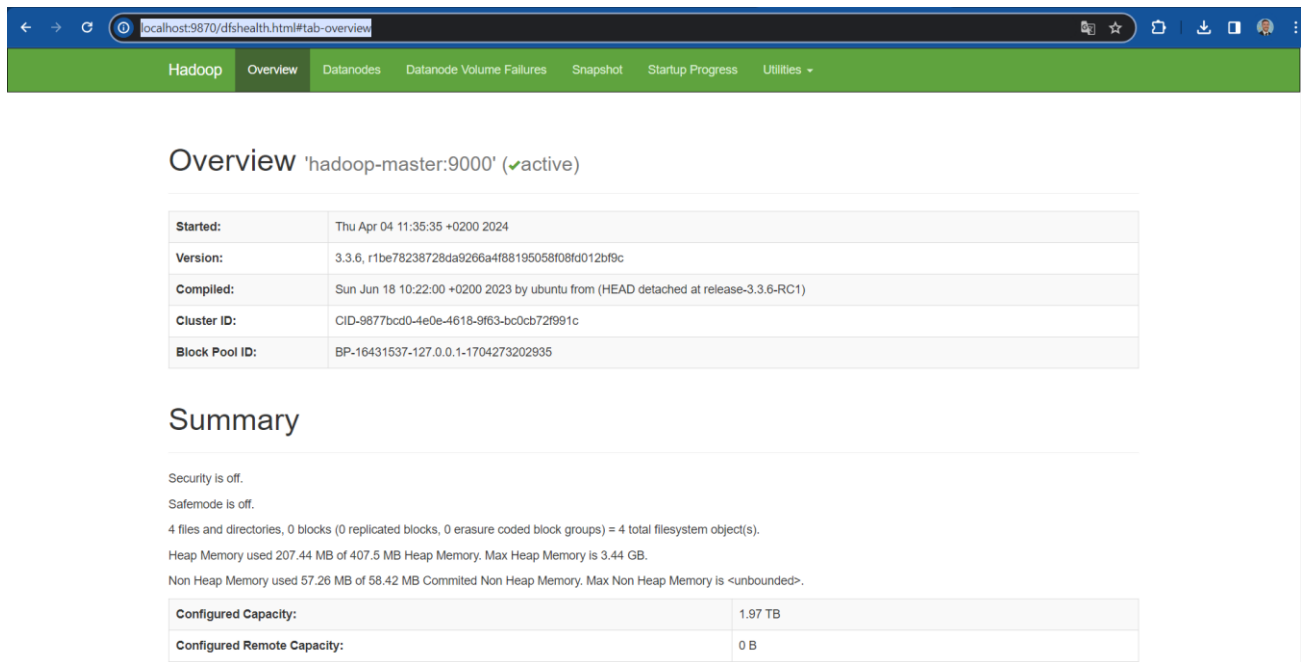
Hadoop:

Framework open source écrit en Java, **hadoop** est utilisé pour le stockage et le traitement distribué (diviser les fichiers en blocs et les réplique) de grandes quantités de données.

- **Le Name Node** : Gère l'espace de nommage, suit les emplacements des blocs de données, coordonne la lecture et l'écriture.
- **Secondary Node** : effectue des sauvegardes régulières des métadonnées du Namenode, améliorer les performances du cluster en optimisant et fusionnant les journaux d'opération. Permet la réduction du temps de récupération en cas de panne du Namenode en donnant une copie des métadonnées.
- **Datanode** : stockage des données sur chaque nœud, reçoit les instructions (lire, écrire, supprimer des blocs de données), envoie des signaux de l'état du nœud, chaque cluster possède plusieurs Datanode.
- **Job Tracker** : Gestion et planification des MapReduce sur le cluster. Communique avec le TaskTracker qui effectue les tâches.
- **TaskTracker** : Exécute les tâches MapReduce sur les nœuds individuels du cluster. Réceptionne les tâches assignées par le TaskTracker, les exécute, communique l'avancement.
- **Map (to value key) / Reduce:**
 - **Map** : transforme en clé, valeur
 - **Shuffle** : tri et regroupement
 - **Reduce** : Combinaison des paires de clé, valeur qui ont la même clé

Le distributed Data Storage , permet de pouvoir lire, écrire un fichier.

Présentation de l'interface Hadoop :



Overview 'hadoop-master:9000' (✓active)

Started:	Thu Apr 04 11:35:35 +0200 2024
Version:	3.3.6, r1be78238728da9266a4f88195058f08fd012bf9c
Compiled:	Sun Jun 18 10:22:00 +0200 2023 by ubuntu from (HEAD detached at release-3.3.6-RC1)
Cluster ID:	CID-9877bcd0-4e0e-4618-9f63-bc0cb72f991c
Block Pool ID:	BP-16431537-127.0.0.1-1704273202935

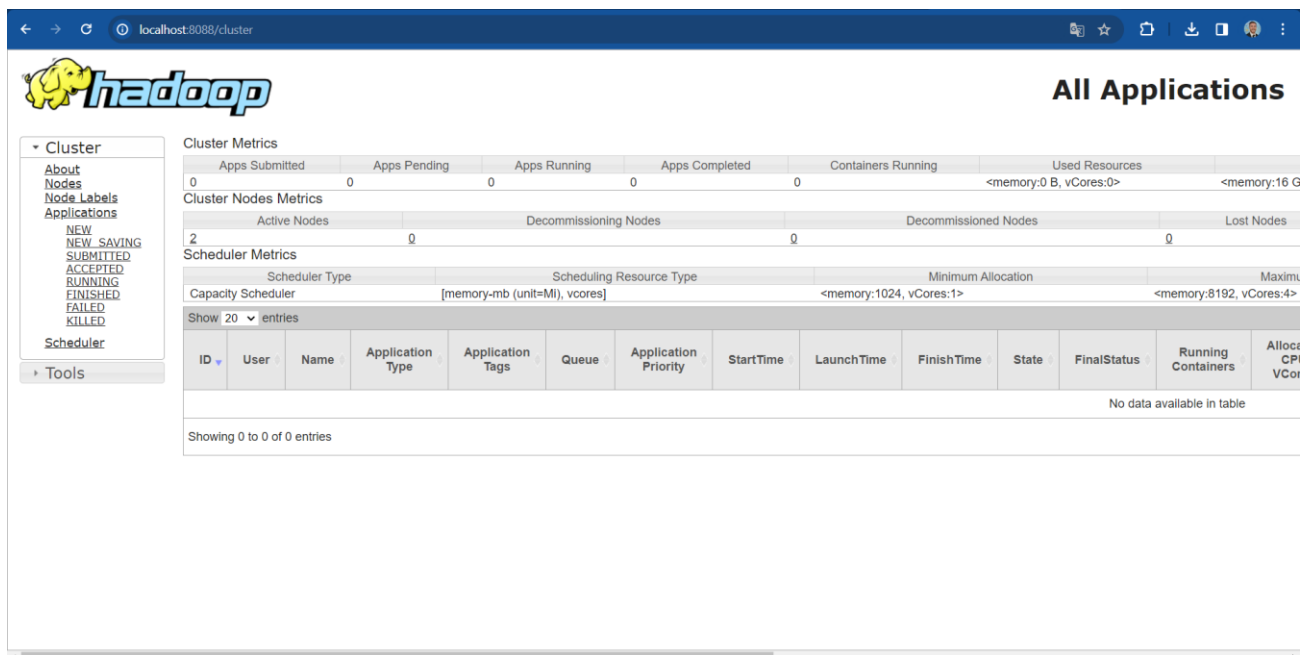
Summary

Security is off.
Safemode is off.

4 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 4 total filesystem object(s).
Heap Memory used 207.44 MB of 407.5 MB Heap Memory. Max Heap Memory is 3.44 GB.
Non Heap Memory used 57.26 MB of 58.42 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	1.97 TB
Configured Remote Capacity:	0 B

Hadoop cluster:



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources
0	0	0	0	0	<memory:0 B, vCores:0> <memory:16 G

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes
2	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCore
No data available in table													

Showing 0 to 0 of 0 entries

Les commandes faites :

Description	Commande
Copier un fichier dans le container docker	<code>docker cp purchases.txt id_container:/purchases.txt</code>
Se reconnecter au container	<code>docker exec -it hadoop-master bash</code>
Placer le fichier présent dans le container docker au dossier input/	<code>hdfs dfs -put purchases.txt input/purchases.txt</code>
Voir la fin du fichier	<code>hdfs dfs -tail input/purchases.txt</code>
Voir le début du fichier	<code>hdfs dfs -head input/purchases.txt</code>
Copier le fichier reduce.py dans le root	<code>docker cp C:\Users\valen\Desktop\reduce.py ae9d0a1b79ce:root/reduce.py</code>
Copier le fichier map.py dans le root	<code>docker cp C:\Users\valen\Desktop\map.py ae9d0a1b79ce:root/map.py</code>
Trouver le fichier mapred-site.xml	<code>find . -name mapred-site.xml</code>
Mettre à jour le node	<code>Apt-get update</code>
Installer Nano	<code>Apt install nano</code>
Modifier le fichier /usr/local/hadoop/etc/hadoop/mapred-site.xml	<code>Cd <<<path>>></code> <code>Nano mapred-site.xml</code> (Code sur teams ou discord)
Permet de lancer le stream d'exécution du map/reduce qui va être décomposé dans les différents workers	<code>hadoop jar \$HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-*.jar -files map.py,reduce.py -mapper "python3 map.py" -reducer "python3 reduce.py" -input /user/root/input/purchases.txt -output /user/root/output1</code>
Récupérer le fichier de HDFS au container	<code>hdfs dfs -get output1/part-00000 /home/</code>
Récupérer le fichier du container au pc en local	<code>docker cp ae9d0a1b79ce:/home/part-00000 C:/Users/valen/Desktop/part-00000</code>

```
PS C:\Users\valen\OneDrive - Efrei\EFREI\EFREI M1\data engineering\projet\TP1BigData> docker cp purchases.txt ae9d0a1b79ce:/purchases.txt
Successfully copied 211MB to ae9d0a1b79ce:/purchases.txt
```

```
root@hadoop-master:/# ll
total 206432
drwxr-xr-x 1 root root 4096 Apr 4 10:20 ./
drwxr-xr-x 1 root root 4096 Apr 4 10:20 ../
-rwxr-xr-x 1 root root 0 Apr 4 09:31 .dockerenv*
lrwxrwxrwx 1 root root 7 Nov 1 2022 bin -> usr/bin/
drwxr-xr-x 2 root root 4096 Apr 18 2022 boot/
drwxr-xr-x 5 root root 360 Apr 4 09:31 dev/
drwxr-xr-x 1 root root 4096 Apr 4 09:31 etc/
drwxr-xr-x 2 root root 4096 Apr 18 2022 home/
lrwxrwxrwx 1 root root 7 Nov 1 2022 lib -> usr/lib/
lrwxrwxrwx 1 root root 9 Nov 1 2022 lib32 -> usr/lib32/
lrwxrwxrwx 1 root root 9 Nov 1 2022 lib64 -> usr/lib64/
lrwxrwxrwx 1 root root 10 Nov 1 2022 libx32 -> usr/libx32/
drwxr-xr-x 2 root root 4096 Nov 1 2022 media/
drwxr-xr-x 2 root root 4096 Nov 1 2022 mnt/
drwxr-xr-x 2 root root 4096 Nov 1 2022 opt/
dr-xr-xr-x 313 root root 0 Apr 4 09:31 proc/
-rwxr-xr-x 1 root root 211312924 Sep 9 2013 purchases.txt*
drwx----- 1 root root 4096 Apr 4 10:18 root/
drwxr-xr-x 1 root root 4096 Apr 4 09:35 run/
lrwxrwxrwx 1 root root 8 Nov 1 2022 sbin -> usr/sbin/
drwxr-xr-x 2 root root 4096 Nov 1 2022 srv/
dr-xr-xr-x 11 root root 0 Apr 4 09:31 sys/
drwxrwxrwt 1 root root 4096 Apr 4 09:35 tmp/
drwxr-xr-x 1 root root 4096 Nov 1 2022 usr/
drwxr-xr-x 1 root root 4096 Nov 1 2022 var/
```

```
root@hadoop-master:/# hdfs dfs -ls input/purchases.txt
-rw-r--r-- 2 root supergroup 211312924 2024-04-04 10:25 input/purchases.txt
```

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

```
2012-0root@hadoop-master:/# hdfs dfs -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
```

```
Click here to ask Blackbox to help you code faster
1 fichier = open("./purchases.txt")
2
3 occurrences = {}
4
5 for ligne in fichier:
6     ligne = ligne.strip()
7     ligne = ligne.lower()
8
9     mots = ligne.split( )
10
11     for mot in mots:
12         if mot.isalpha():
13             if mot in occurrences:
14                 occurrences[mot] += 1
15             else:
16                 occurrences[mot] = 1
17
18 for key in list(occurrences.keys()):
19     print(key+": "+str(occurrences[key]))
```

localhost:9870/explorer.html#/user/root/input

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Browse Directory

/user/root/input

Show 25 entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	-rw-r--r--	root	supergroup	201.52 MB	Apr 04 12:25	2	128 MB	purchases.txt	<input type="button" value="Delete"/>

Showing 1 to 1 of 1 entries

Previous 1 Next