TP 2 Big Data

Exécution du programme en local :

```
PS C:\Users\21656\Desktop\ING 2\big-data\tp\tp2> cat file.txt |python mapper.py | sort |python reducer.py
Big     2
Bonjour 2
coeur   1
Data    2
du      1
est     1
et      1
Hadoop  2
le      1
Spark   1
PS C:\Users\21656\Desktop\ING 2\big-data\tp\tp2>
```

Exécution du programme sur le cluster :

1. Transférer les scripts dans le cluster Hadoop :

```
PS C:\Users\21656\Desktop\ING 2\big-data\tp\tp2> docker cp mapper.py namenode:/mapperWC.py
Successfully copied 2.05kB to namenode:/mapperWC.py
PS C:\Users\21656\Desktop\ING 2\big-data\tp\tp2> docker cp reducer.py namenode:/reducerWC.py
Successfully copied 2.56kB to namenode:/reducerWC.py
PS C:\Users\21656\Desktop\ING 2\big-data\tp\tp2>
```

2. Entrer dans le container du namenode et tester les scripts :

```
PS C:\Users\21656\Desktop\ING 2\big-data\tp\tp2> docker exec -it namenode bash
root@c83a2ab504b2:/# chmod u+x mapperWC.py
root@c83a2ab504b2:/# chmod u+x reducerWC.py
root@c83a2ab504b2:/# cat input.txt |python mapperWC.py | sort |python reducerWC.py
```

Tester les scripts sur le container:

```
root@c83a2ab504b2:/# cat input.txt |python mapperWC.py | sort |python reducerWC.py
Big     2
Bonjour 2
Data    2
Hadoop  2
Spark   1
coeur   1
du      1
est     1
et      1
le      1
root@c83a2ab504b2:/#
```

3. Créer des fichiers textes et les transférer dans HDFS :

```
root@c83a2ab504b2:/# mkdir input
root@c83a2ab504b2:/# echo "Hello World" > input/f1.txt
root@c83a2ab504b2:/# echo "Hello Docker" > input/f2.txt
root@c83a2ab504b2:/# echo "Hello Hadoop" > input/f3.txt
root@c83a2ab504b2:/# echo "Hello MapReduce" > input/f4.txt
root@c83a2ab504b2:/# hadoop fs -mkdir -p input
root@c83a2ab504b2:/# hdfs dfs -put  ./input/*  input
2024-10-25 13:40:21,941 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
2024-10-25 13:40:22,077 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
2024-10-25 13:40:22,547 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
2024-10-25 13:40:22,597 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
root@c83a2ab504b2:/#
```

4. Exécuter le programme MapReduce:

```
root@c83a2ab504b2:/#find / -name 'hadoop-streaming*.jar'
/opt/hadoop-3.2.1/share/hadoop/tools/lib/hadoop-streaming-3.2.1.jar
/opt/hadoop-3.2.1/share/hadoop/tools/sources/hadoop-streaming-3.2.1-test-sources.jar
/opt/hadoop-3.2.1/share/hadoop/tools/sources/hadoop-streaming-3.2.1-sources.jar
root@c83a2ab504b2:/#
```

```
root@c83a2ab504b2:/# hadoop jar /opt/hadoop-3.2.1/share/hadoop/tools/lib/hadoop-streaming-3.2.1.jar -files mapperWC.py,reducerWC.py -input input -output out
put -mapper "python3 mapperWC.py" -reducer "python3 reducerWC.py"
packageJobJar: [/tmp/hadoop-unjar7850044465117329043/] [] /tmp/streamjob6158240412648556696.jar tmpDir=null
2024-10-25 14:09:34,019 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/172.18.0.6:8032
2024-10-25 14:09:34,173 INFO client.AHSProxy: Connecting to Application History server at historyserver/172.18.0.3:10200
2024-10-25 14:09:34,199 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/172.18.0.6:8032
2024-10-25 14:09:34,200 INFO client.AHSProxy: Connecting to Application History server at historyserver/172.18.0.3:10200
2024-10-25 14:09:34,393 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1729862204996_0004
2024-10-25 14:09:34,512 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-10-25 14:09:34,613 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-10-25 14:09:34,655 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-10-25 14:09:35,196 INFO mapred.FileInputFormat: Total input files to process : 4
2024-10-25 14:09:35,247 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-10-25 14:09:35,718 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-10-25 14:09:36,154 INFO mapreduce.JobSubmitter: number of splits:4
2024-10-25 14:09:36,268 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-10-25 14:09:36,757 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1729862204996_0004
2024-10-25 14:09:36,757 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-10-25 14:09:36,920 INFO conf.Configuration: resource-types.xml not found
2024-10-25 14:09:36,922 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-10-25 14:09:37,185 INFO impl.YarnClientImpl: Submitted application application_1729862204996_0004
2024-10-25 14:09:37,214 INFO mapreduce.Job: The url to track the job: http://resourcemanager:8088/proxy/application_1729862204996_0004/
2024-10-25 14:09:37,217 INFO mapreduce.Job: Running job: job_1729862204996_0004
2024-10-25 14:09:42,429 INFO mapreduce.Job: Job job_1729862204996_0004 running in uber mode : false
2024-10-25 14:09:42,431 INFO mapreduce.Job:  map 0% reduce 0%
2024-10-25 14:09:47,503 INFO mapreduce.Job:  map 25% reduce 0%
2024-10-25 14:09:49,524 INFO mapreduce.Job:  map 50% reduce 0%
2024-10-25 14:09:50,539 INFO mapreduce.Job:  map 75% reduce 0%
2024-10-25 14:09:51,553 INFO mapreduce.Job:  map 100% reduce 0%
2024-10-25 14:09:53,569 INFO mapreduce.Job:  map 100% reduce 100%
2024-10-25 14:09:54,595 INFO mapreduce.Job: Job job_1729862204996_0004 completed successfully
2024-10-25 14:09:54,662 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=75
                FILE: Number of bytes written=1164327
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=434
                HDFS: Number of bytes written=46
                HDFS: Number of read operations=17
```

```
2024-10-25 14:09:54,662 INFO streaming.StreamJob: Output directory: output
```

5. Voir les résultats :

```
root@c83a2ab504b2:/# hdfs dfs -ls output
Found 2 items
-rw-r--r--   3 root supergroup          0 2024-10-25 14:09 output/_SUCCESS
-rw-r--r--   3 root supergroup         46 2024-10-25 14:09 output/part-00000
root@c83a2ab504b2:/#
```

```
root@c83a2ab504b2:/# hdfs dfs -cat output/part-00000
2024-10-25 14:14:34,821 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Docker  1
Hadoop  1
Hello   4
MapReduce       1
World   1
```