

# DÉPARTEMENT MATHÉMATIQUE INFORMATIQUE

Big Data : Fondements et Architectures de stockage

## Rapport

### TP 1 : MANIPULATION DU SYSTÈME DE FICHIERS HDFS

Réaliser par :  
ETOULLALI Ayoub

Professeur :  
Mr. BOUSSELHAM Abdelmajid

2ème année II-BDCC

Filière d'ingénieur : Ingénieur informatique, Big Data et Cloud Computing

# SOMMAIRE

<b>Introduction .....</b>	<b>2</b>
<b>Démarrez les processus Hadoop .....</b>	<b>3</b>
<b>Vérifiez l'exécution .....</b>	<b>4</b>
<b>Accédez à l'interface web de NameNode .....</b>	<b>4</b>
<b>Créez l'arborescence dans la racine du HDFS .....</b>	<b>4</b>
<b>Afficher le contenu des fichiers .....</b>	<b>5</b>
<b>Copiez des fichiers .....</b>	<b>5</b>
<b>Supprimez un fichier et renommez des fichiers .....</b>	<b>6</b>
<b>Copier les fichiers à partir du système de fichier local vers le répertoire TPs .....</b>	<b>7</b>
<b>Conclusion .....</b>	<b>8</b>



Ce rapport présente les résultats du TP1 du cours "Big Data : Fondements et Architectures de stockage".

Ce TP est consacré à la manipulation du système de fichiers HDFS et a pour but de la familiarisation avec les commandes Hadoop pour créer des répertoires, des fichiers, copier, déplacer et supprimer des fichiers dans le HDFS. Les tâches incluent également la création d'arborescences de répertoires, la copie de fichiers à partir du système de fichiers local et l'affichage du contenu des fichiers. Les objectifs de ce TP sont d'acquérir une compréhension pratique du *Hadoop Distributed File System* et de maîtriser les commandes Hadoop pour gérer les fichiers dans un environnement Big Data.

Les commandes pour accomplir les tâches décrites dans le TP sont les suivantes :

## Démarrez les processus Hadoop

```
PS C:\Windows\System32> hdfs namenode -format
```

```
PowerShell 7.3.2
PS C:\Windows\System32> hdfs namenode -format
23/02/13 15:07:22 INFO namenode.NameNode: STARTUP_MSG:
/*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = ACER/192.168.56.1
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 2.7.3
```

```
PS C:\Windows\System32> start-dfs.cmd
```

```
Apache Hadoop Distribution - hadoop namenode
23/02/13 15:08:27 INFO namenode.FSImage: No edit log streams selected.
23/02/13 15:08:27 INFO namenode.FSImage: Planning to load image: FSImageFile(file=C:\app\hadoop\tmp\dfs\name\current\fsimage_00000000000000000000, cpktxid=00000000000000000000)
23/02/13 15:08:27 INFO namenode.FSImageFormatPSIIndex: Loading 1 Indexes.
23/02/13 15:08:27 INFO namenode.FSImageFormatProtobuf: Loaded FSImage in 0 seconds.
23/02/13 15:08:27 INFO namenode.FSImage: Loaded image for tail 0 from C:\app\hadoop\tmp\dfs\name\current\fsimage_00000000000000000000
23/02/13 15:08:27 INFO namenode.FSNamesystem: Need to save fs image? false (stateImage=false, hdfsAbled=false, isRollingUpgrade=False)
23/02/13 15:08:27 INFO namenode.FSImage: Finished loading FSImage in 876 msec
23/02/13 15:08:29 INFO namenode.NameNode: RPC server is binding to localhost:19000
23/02/13 15:08:29 INFO ipc.CallQueueManager: Using callQueue class java.util.concurrent.LinkedBlockingQueue
23/02/13 15:08:29 INFO ipc.Server: Starting Socket Reader #1 for port 19000
23/02/13 15:08:29 INFO namenode.FSNamesystem: Registered FSNamesystemState MBean
23/02/13 15:08:29 WARN common.Util: Path /app/hadoop/tmp/dfs/name should be specified as a URI in configuration files. Please update hdfs configuration.
23/02/13 15:08:29 INFO namenode.LeaseManager: Number of blocks under construction: 0
23/02/13 15:08:29 INFO namenode.LeaseManager: Number of blocks under construction: 0
23/02/13 15:08:29 INFO namenode.FSNamesystem: Initializing replication queues
23/02/13 15:08:29 INFO hdfs.StateChange: STATE* Leaving safe mode after 5 secs
23/02/13 15:08:29 INFO hdfs.StateChange: STATE* Network topology has 0 racks and 0 datanodes
23/02/13 15:08:29 INFO hdfs.StateChange: STATE* UnderReplicatedBlocks has 0 blocks
23/02/13 15:08:29 INFO BlockManagement.DataNodeDescriptor: Number of failed storage changes from 0 to 0
23/02/13 15:08:29 INFO BlockManagement.BlockManager: Total number of blocks = 0
23/02/13 15:08:29 INFO BlockManagement.BlockManager: Number of invalid blocks = 0
23/02/13 15:08:29 INFO BlockManagement.BlockManager: Number of under-replicated blocks = 0
23/02/13 15:08:29 INFO BlockManagement.BlockManager: Number of over-replicated blocks = 0
23/02/13 15:08:29 INFO BlockManagement.BlockManager: Number of blocks being written = 0
23/02/13 15:08:29 INFO hdfs.StateChange: STATE* Replication Queue initialization scan for invalid, over- and under-replicated blocks completed in 208 msec
23/02/13 15:08:29 INFO ipc.Server: IPC Server listener on 19000: starting
23/02/13 15:08:29 INFO ipc.Server: IPC Server Responder: starting
23/02/13 15:08:29 INFO namenode.NameNode: NameNode RPC up at: localhost/127.0.0.1:19000
23/02/13 15:08:29 INFO namenode.FSNamesystem: Starting services required for active state
23/02/13 15:08:30 INFO BlockManagement.CacheReplicationMonitor: Starting CacheReplicationMonitor with interval 30000 milliseconds
23/02/13 15:12:23 INFO window.RollingWindowManager: topx size for command getfileinfo is: 1
23/02/13 15:12:23 INFO window.RollingWindowManager: topx size for command * is: 1
23/02/13 15:12:23 INFO window.RollingWindowManager: topx size for command getfileinfo is: 1
23/02/13 15:12:23 INFO window.RollingWindowManager: topx size for command * is: 1
23/02/13 15:12:23 INFO window.RollingWindowManager: topx size for command getfileinfo is: 1
23/02/13 15:12:23 INFO window.RollingWindowManager: topx size for command * is: 1
```

```
PS C:\Windows\System32> start-yarn.cmd
starting yarn daemons
```

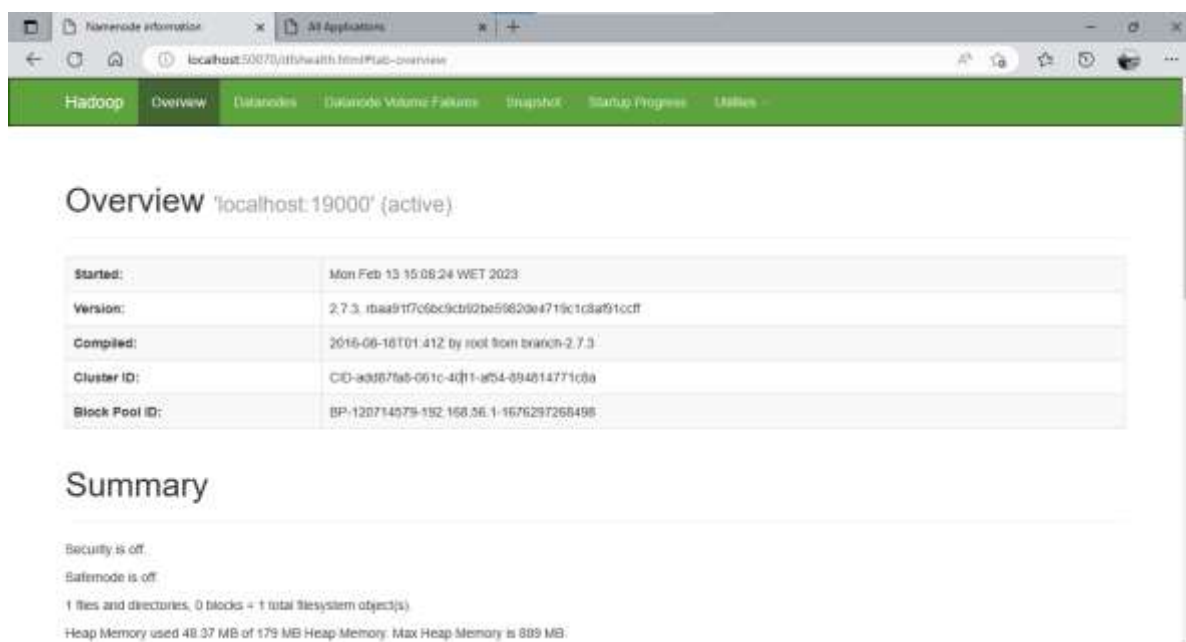
```
Apache Hadoop Distribution - yarn resourcemanager
23/02/13 15:09:07 INFO delegation.AbstractDelegationTokenSecretManager: Starting expired delegation token remover thread, tokenRemoverScanInterval=60 min(s)
23/02/13 15:09:07 INFO delegation.AbstractDelegationTokenSecretManager: Updating the current master key for generating delegation tokens
Füvr. 13, 2023 3:09:08 PM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory register
INFO: Registering org.apache.hadoop.yarn.server.resourcemanager.webapp.JAXContextResolver as a provider class
Füvr. 13, 2023 3:09:08 PM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory register
INFO: Registering org.apache.hadoop.yarn.server.resourcemanager.webapp.RMWebServices as a root resource class
Füvr. 13, 2023 3:09:08 PM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory register
INFO: Registering org.apache.hadoop.yarn.server.resourcemanager.webapp.GenericExceptionHandler as a provider class
Füvr. 13, 2023 3:09:08 PM com.sun.jersey.server.impl.application.WebApplicationImpl _initiate
INFO: Initiating Jersey application, version 'Jersey: 1.9.0/02/2011 11:17 AM'
Füvr. 13, 2023 3:09:08 PM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory getComponentProvider
INFO: Binding org.apache.hadoop.yarn.server.resourcemanager.webapp.JAXContextResolver to GuiceManagedComponentProvider with the scope "Singleton"
Füvr. 13, 2023 3:09:10 PM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory getComponentProvider
INFO: Binding org.apache.hadoop.yarn.webapp.GenericExceptionHandler to GuiceManagedComponentProvider with the scope "Singleton"
Füvr. 13, 2023 3:09:14 PM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory getComponentProvider
INFO: Binding org.apache.hadoop.yarn.server.resourcemanager.webapp.RMWebServices to GuiceManagedComponentProvider with the scope "Singleton"
23/02/13 15:09:15 INFO moribay.log: Started HttpServer2$SelectChannelConnectorWithSafeStartup@0.0.0.0:8088
23/02/13 15:09:15 INFO webapp.WebApp: Web app cluster started at 8088
23/02/13 15:09:15 INFO ipc.CallQueueManager: Using callQueue class java.util.concurrent.LinkedBlockingQueue
23/02/13 15:09:15 INFO pb.RpcServerFactoryPBImpl: Adding protocol org.apache.hadoop.yarn.server.api.ResourceManagerAdministrationProtocolPB to the server
23/02/13 15:09:15 INFO ipc.Server: IPC Server Responder: starting
23/02/13 15:09:15 INFO ipc.Server: Starting Socket Reader #1 for port 8033
23/02/13 15:09:15 INFO ipc.Server: IPC Server listener on 8033: starting
23/02/13 15:09:18 INFO util.RackResolver: Resolved ACER to /default-rack
23/02/13 15:09:18 INFO resourcemanager.ResourceTrackerService: NodeManager from node ACER(cmpPort: 58636 httpPort: 8042) registered with capability: <memory:10000, vCores:8>, assigned nodeId ACER:58636
23/02/13 15:09:18 INFO rmnode.RMNodeImpl: ACER:58636 Node Transitioned from NEW to RUNNING
23/02/13 15:09:18 INFO capacity.CapacityScheduler: Added node ACER:58636 clusterResource: <memory:10000, vCores:8>
23/02/13 15:09:18 INFO rmnode.RMNodeImpl: Node ACER:58636 reported UNHEALTHY with details: 1/1 local-dirs are bad: /tmp/hadoop-pc/rm-local-dir; 1/1 log-dirs are bad: C:\hadoop-2.7.3\logs\userlogs
23/02/13 15:09:18 INFO rmnode.RMNodeImpl: ACER:58636 Node Transitioned from RUNNING to UNHEALTHY
23/02/13 15:09:18 INFO capacity.CapacityScheduler: Removed node ACER:58636 clusterResource: <memory:0, vCores:0>
23/02/13 15:19:00 INFO scheduler.AbstractYarnScheduler: Release request cache is cleaned up
```

## Vérifiez l'exécution

```
PS C:\Users\pc> jps
9696 NodeManager
6452 NameNode
13272 DataNode
2856 Jps
3436 ResourceManager
```

## Accédez à l'interface web de NameNode

Ouvrez un navigateur web et entrez l'URL suivant : <http://localhost:50070>



## Créez l'arborescence dans la racine du HDFS

```
PS C:\Windows\System32> hdfs dfs -mkdir /BDCC
```

```
PS C:\Windows\System32> hdfs dfs -mkdir /BDCC/JAVA
PS C:\Windows\System32> hdfs dfs -mkdir /BDCC/CPP
PS C:\Windows\System32> hdfs dfs -mkdir /BDCC/CPP/TPs
PS C:\Windows\System32> hdfs dfs -mkdir /BDCC/CPP/Cours
PS C:\Windows\System32> hdfs dfs -mkdir /BDCC/JAVA/Cours
PS C:\Windows\System32> hdfs dfs -mkdir /BDCC/JAVA/TPs
```

La commande <<echo>> ne fonctionne pas, donc j'ai fait une copie de l'ordinateur local

```
PS C:\Windows\System32> hdfs dfs -copyFromLocal C:\Users\pc\Desktop\ENSET\S4\CoursCPP1.txt /BDCC/CPP/Cours
PS C:\Windows\System32> hdfs dfs -copyFromLocal C:\Users\pc\Desktop\ENSET\S4\CoursCPP2.txt /BDCC/CPP/Cours
PS C:\Windows\System32> hdfs dfs -copyFromLocal C:\Users\pc\Desktop\ENSET\S4\CoursCPP3.txt /BDCC/CPP/Cours
```

## Afficher le contenu des fichiers

```
PS C:\Windows\System32> hdfs dfs -cat /BDDC/PHP/Cours/CoursCPP1.txt
Contenu 1
PS C:\Windows\System32> hdfs dfs -cat /BDDC/PHP/Cours/CoursCPP2.txt
Contenu 2
PS C:\Windows\System32> hdfs dfs -cat /BDDC/PHP/Cours/CoursCPP3.txt
Contenu 3
```

File information - CoursCPP1.txt

✕

[Download](#)

Block information —

Block 0 ▾

Block ID: 1073741828

Block Pool ID: BP-2133203550-10.10.42.18-1676911286819

Generation Stamp: 1005

Size: 11

Availability:

- ACER.mshome.net

## Copiez des fichiers

```
PS C:\Windows\System32> hdfs dfs -cp /BDDC/PHP/Cours/CoursCPP3.txt /BDDC/JAVA/Cours
PS C:\Windows\System32> hdfs dfs -cp /BDDC/PHP/Cours/CoursCPP2.txt /BDDC/JAVA/Cours
PS C:\Windows\System32> hdfs dfs -cp /BDDC/PHP/Cours/CoursCPP1.txt /BDDC/JAVA/Cours
```

/BDDC/PHP/Cours

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	pc	supergroup	11 B	20/02/2023 21:53:43	1	128 MB	<a href="#">CoursCPP1.txt</a>
-rw-r--r--	pc	supergroup	11 B	20/02/2023 21:53:59	1	128 MB	<a href="#">CoursCPP2.txt</a>
-rw-r--r--	pc	supergroup	11 B	20/02/2023 21:54:15	1	128 MB	<a href="#">CoursCPP3.txt</a>



## Supprimez un fichier et renommez des fichiers

```
PS C:\Windows\System32> hdfs dfs -rm /BDDC/CPP/Cours/CoursCPP3.txt
23/02/20 22:04:50 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /BDDC/CPP/Cours/CoursCPP3.txt
```

/BDDC/CPP/Cours

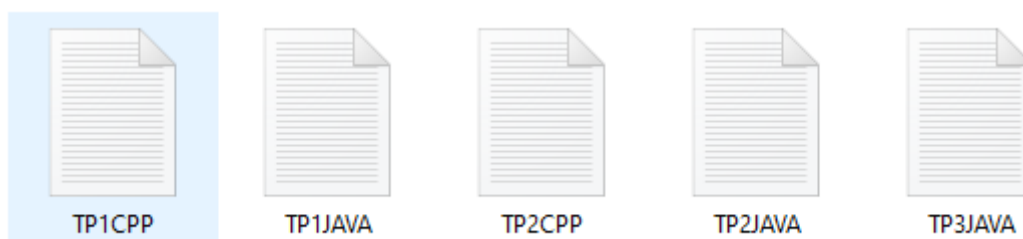
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	pc	supergroup	11 B	20/02/2023 21:53:43	1	128 MB	<a href="#">CoursCPP1.txt</a>
-rw-r--r--	pc	supergroup	11 B	20/02/2023 21:53:59	1	128 MB	<a href="#">CoursCPP2.txt</a>

```
PS C:\Users\pc> hdfs dfs -mv /BDDC/CPP/Cours/CoursCPP2.txt /BDDC/JAVA/Cours/CoursJAVA2.txt
PS C:\Users\pc> hdfs dfs -mv /BDDC/CPP/Cours/CoursCPP1.txt /BDDC/JAVA/Cours/CoursJAVA1.txt
PS C:\Users\pc> hdfs dfs -rm /BDDC/JAVA/Cours/CoursCPP1.txt
23/02/20 22:30:51 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /BDDC/JAVA/Cours/CoursCPP1.txt
PS C:\Users\pc> hdfs dfs -rm /BDDC/JAVA/Cours/CoursCPP2.txt
23/02/20 22:31:00 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /BDDC/JAVA/Cours/CoursCPP2.txt
```

/BDDC/JAVA/Cours

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	pc	supergroup	11 B	20/02/2023 22:27:01	1	128 MB	<a href="#">CoursJAVA1.txt</a>
-rw-r--r--	pc	supergroup	11 B	20/02/2023 22:27:13	1	128 MB	<a href="#">CoursJAVA2.txt</a>

## Copier les fichiers à partir du système de fichier local vers le répertoire TPs



```
PS C:\Users\pc> hdfs dfs -copyFromLocal C:\Users\pc\Desktop\ENSET\S4\Big data\TPs\TP1\TP1JAVA.txt /BDDC/JAVA/TPs
PS C:\Users\pc> hdfs dfs -copyFromLocal C:\Users\pc\Desktop\ENSET\S4\Big data\TPs\TP1\TP2JAVA.txt /BDDC/JAVA/TPs
PS C:\Users\pc> hdfs dfs -copyFromLocal C:\Users\pc\Desktop\ENSET\S4\Big data\TPs\TP1\TP1CPP.txt /BDDC/CPP/TPs
PS C:\Users\pc> hdfs dfs -copyFromLocal C:\Users\pc\Desktop\ENSET\S4\Big data\TPs\TP1\TP2CPP.txt /BDDC/CPP/TPs
PS C:\Users\pc> hdfs dfs -copyFromLocal C:\Users\pc\Desktop\ENSET\S4\Big data\TPs\TP1\TP3JAVA.txt /BDDC/JAVA/TPs
```

/BDDC/JAVA/TPs							
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	pc	supergroup	0 B	20/02/2023 22:37:07	1	128 MB	TP1JAVA.txt
-rw-r--r--	pc	supergroup	0 B	20/02/2023 22:37:39	1	128 MB	TP2JAVA.txt
-rw-r--r--	pc	supergroup	0 B	20/02/2023 22:38:37	1	128 MB	TP3JAVA.txt

/BDDC/CPP/TPs							
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	pc	supergroup	0 B	20/02/2023 22:37:58	1	128 MB	TP1CPP.txt
-rw-r--r--	pc	supergroup	0 B	20/02/2023 22:38:23	1	128 MB	TP2CPP.txt



## Conclusion

En conclusion, ce TP a été très instructif et nous a permis de mieux comprendre le fonctionnement du Hadoop Distributed File System et les commandes Hadoop pour gérer les fichiers dans un environnement Big Data. Nous avons appris comment créer des répertoires et des fichiers, ajouter du contenu, copier, déplacer et supprimer des fichiers dans le HDFS. Nous avons également appris comment créer des arborescences de répertoires, copier des fichiers à partir du système de fichiers local et afficher le contenu des fichiers. Nous sommes désormais mieux équipés pour gérer des fichiers dans un environnement Big Data et nous avons hâte de poursuivre notre apprentissage dans ce domaine passionnant.



**Ayoub ETOULLALI**  
Élève ingénieur en Ingénierie Informatique  
Big Data & Cloud Computing  
Université Hassan II de Casablanca | ENSET Mohammeda

+212 6 58 71 20 11  
ayoub.etoullali2002@gmail.com  
<https://github.com/Ayoub-etoullali>  
ERRACHIDIA

