

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA
RECHERCHE SCIENTIFIQUE

UNIVERSITE BADJI MOKHTAR ANNABA
FACULTE DES SCIENCES DE L'INGENIORAT
DEPARTEMENT INFORMATIQUE

HADOOP MAPREDUCE

Master : Gestion et Analyse des Données Massives (GADM)

2^{me} année

Dr. Klai Sihem

TABLE DES MATIÈRES

PRÉFACE	3
TABLE DES MATIÈRES	6
LISTE DES FIGURES	7
A ANNEXES	1
A.1 TPN° 1 : L'INSTALLATION ET LA VÉRIFICATION D'UN ENSEMBLE DE PRÉ- REQUIS	2
A.1.1 Installation de Cygwin	2
A.1.2 Définition des variables d'environnement	3
A.1.3 Installer le service SSH	5
A.2 TPN° 2 : INSTALLATION DE HADOOP	8
A.3 TPN° 3 : SE FAMILIARISER AVEC LES COMMANDES HDFS	12
A.4 TPN° 4 : IMPLÉMENTER UN PROGRAMME MAPREDUCE : EXERCICE1 . . .	15
A.5 TPN° 5 : IMPLÉMENTER LE PROGRAMME MAPREDUCE DE L'EXERCICE2 . .	19
A.6 TPN° 6 : IMPLÉMENTER LE PROGRAMME MAPREDUCE DE L'EXERCICE3 : QUESTION2	22

LISTE DES FIGURES

A.1	Installation de Cygwin	3
A.2	Installation de openssh	3
A.3	Modification de la variable path	4
A.4	Modification de la variable path (suite)	4
A.5	Configuration du service ssh	5
A.6	Démarrer le service ssh	6
A.7	Configuration de la clé d'autorization	7
A.8	Copier le fichier archive Hadoop	8
A.9	Copier le fichier archive Hadoop	9
A.10	le fichier <code>etc\hadoop\core-site.xml</code>	10
A.11	le fichier <code>etc\hadoop\hdfs-site.xml</code>	11
A.12	le fichier <code>etc\hadoop\mapred-site.xml</code>	11
A.13	Formater HDFS	11
A.14	Commande hdfs	12
A.15	Commande hdfs dfs	13
A.16	Commande hdfs dfs -cat appel-T.txt	13
A.17	Commande hdfs dfs -put *.txt entree	14
A.18	Commande hdfs dfs -rm in.txt	14
A.19	Commande hdfs dfs -mkdir entree	15
A.20	Insertion des fichiers Jar Hadoop dans le projet	16
A.21	Insertion du programme de l'exercice1 du chapitre ?? dans Eclipse	17
A.22	Génération du fichier JAR	17
A.23	Vérification si le programme .jar généré existe	17
A.24	Exécution du programme	18

A.25	Affichage des résultats	18
A.26	Apperçu du programme dans Eclipse : Exercice 2 Question1 . . .	19
A.27	Lancement du programme Exercice 2 Question1	19
A.28	Affichage du résultat Exercice 2 Question1	20
A.29	Apperçu du programme dans Eclipse : Exercice 2 Question2 . . .	20
A.30	Lancement du programme Exercice 2 Question2 et affichage du contenu du fichier in.txt	21
A.31	Affichage du résultat Exercice 2 Question2	21
A.32	Affichage des fichiers en entrée : Produit et Vente	22
A.33	Exécution du programme	23
A.34	Affichage du résultat	23

ANNEXES

A

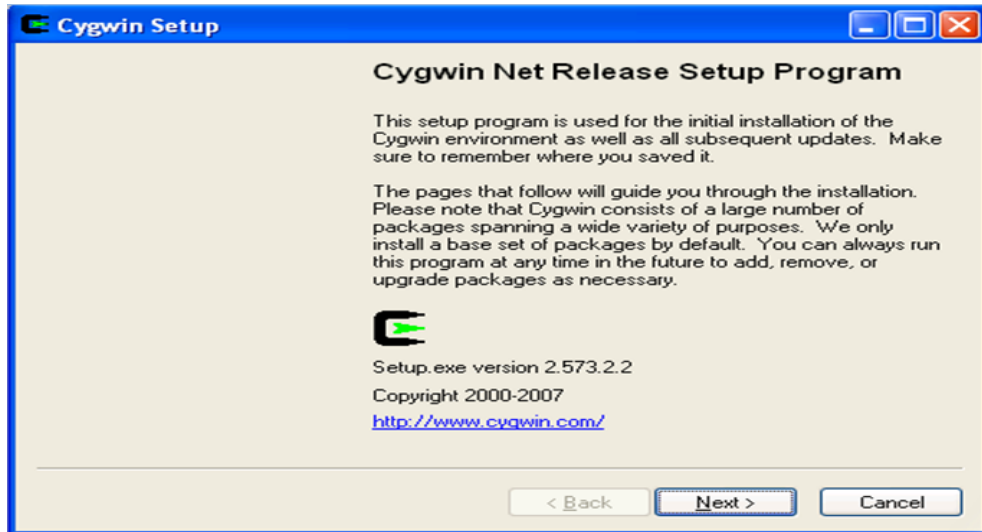
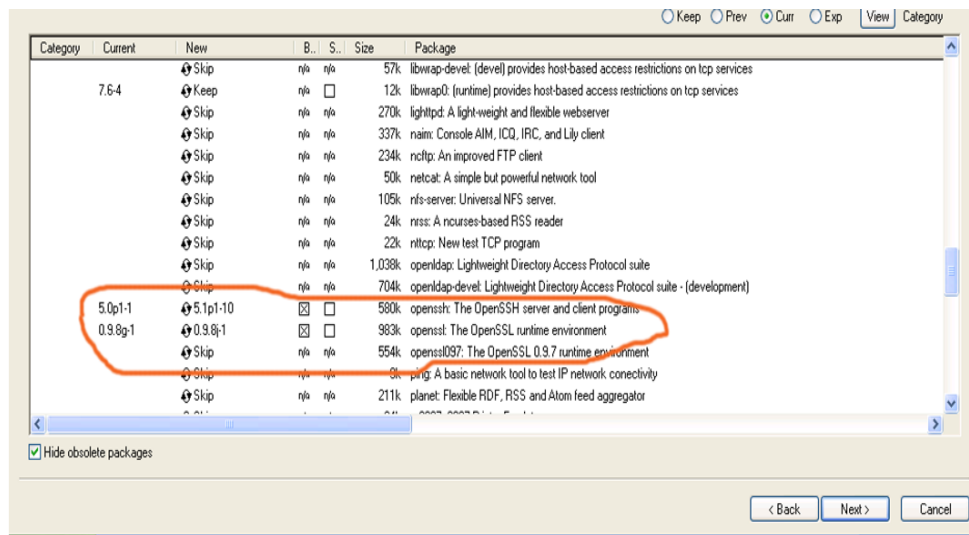
A.1 TPN°1 : L'INSTALLATION ET LA VÉRIFICATION D'UN ENSEMBLE DE PRÉ-REQUIS

Une version récente du logiciel Java doit être installée dans le disque C:\

A.1.1 Installation de Cygwin

Cygwin est un ensemble de packages de Unix portés sur Microsoft Windows. Il est nécessaire pour exécuter Hadoop sous windows car Hadoop est écrit pour la plate-forme Unix. Voici les étapes à suivre pour installer Cygwin :

1. Télécharger Cygwin à partir de <http://www.cygwin.com>.
2. Executer le fichier téléchargé ;
3. Si vous avez l'écran A.1, tapez le bouton Next ;
4. Puis vous aurez la liste des packages affichée, figure A.2. Cliquez sur [openssh](#) , c'est le package nécessaire pour le fonctionnement de Hadoop ;
5. Puis tapez Next pour completer l'installation.

FIGURE A.1 – *Installation de Cygwin*FIGURE A.2 – *Installation de openssl*

A.1.2 Définition des variables d'environnement

Il est nécessaire de modifier la variable PATH. Pour cela, suivre ces étapes :

1. Tapez **Panneau de configuration / système et sécurité / système / paramètre système avancés** ;
2. Quand cette boîte de dialogue **A.3** apparaît cliquez sur **Environment Variables / Path / Edit A.4** ;
3. Dans la zone **variable** , ajoutez le chemin de Cygwin voir exemple : **c:\cygwin\bin;c:\cygwin\usr\bin**
4. Fermez les boîtes de dialogues.

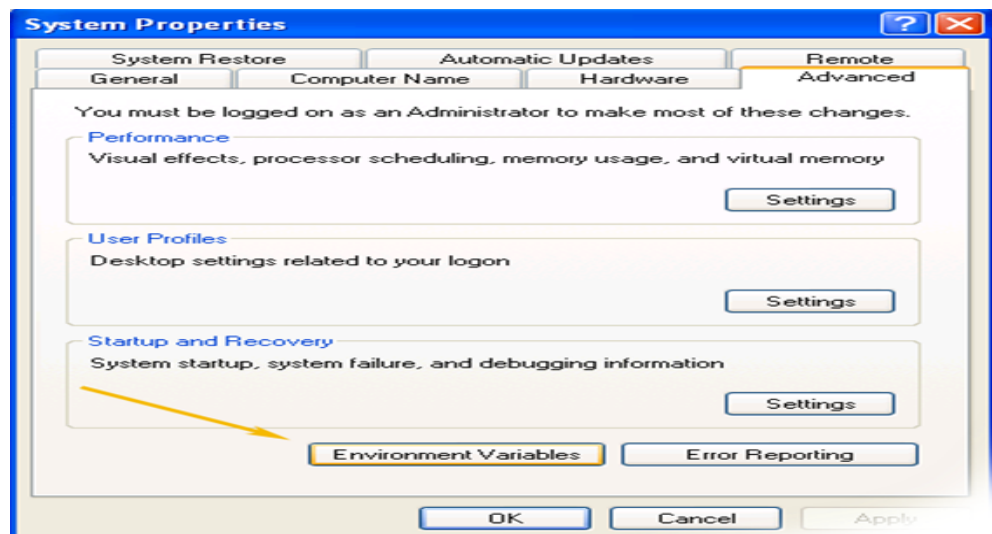


FIGURE A.3 – Modification de la variable path

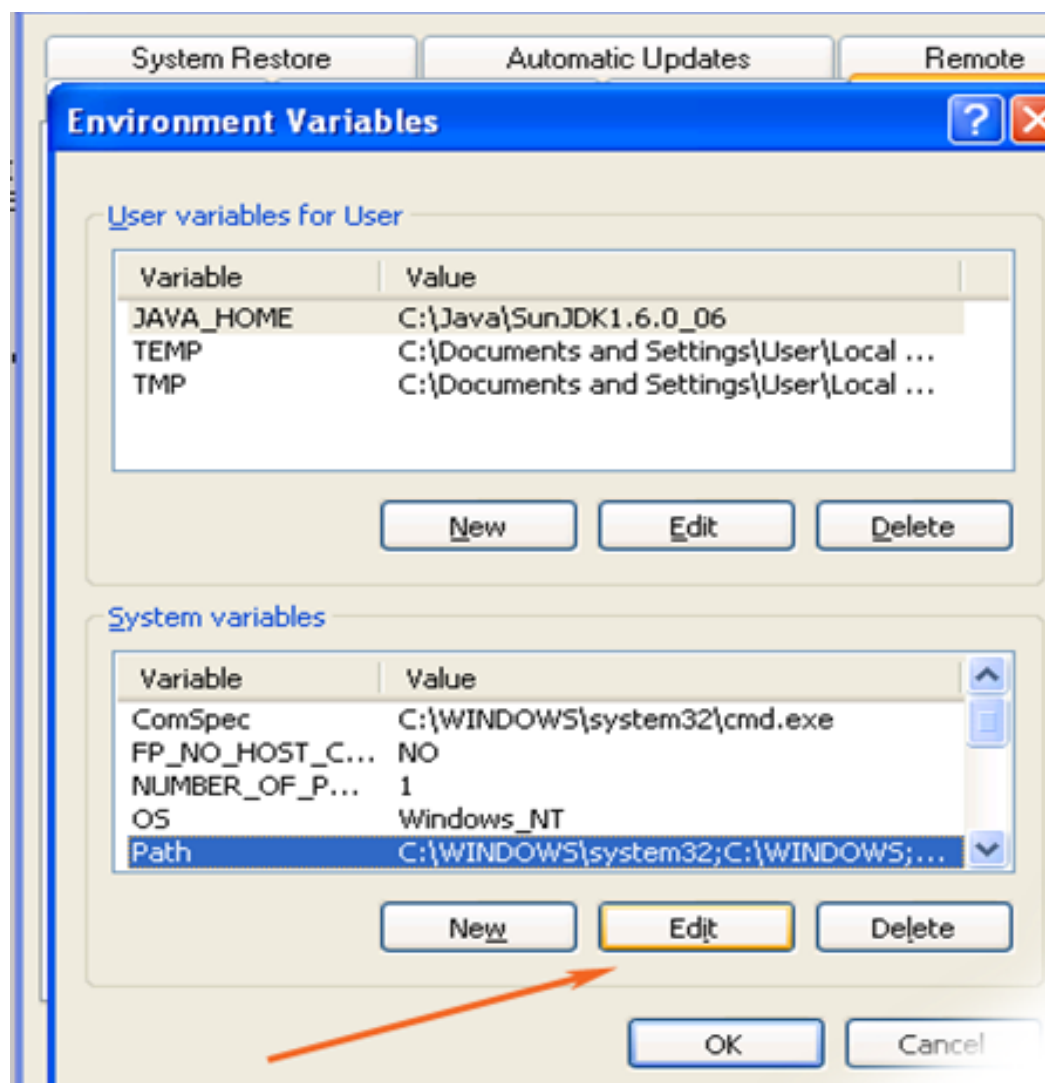


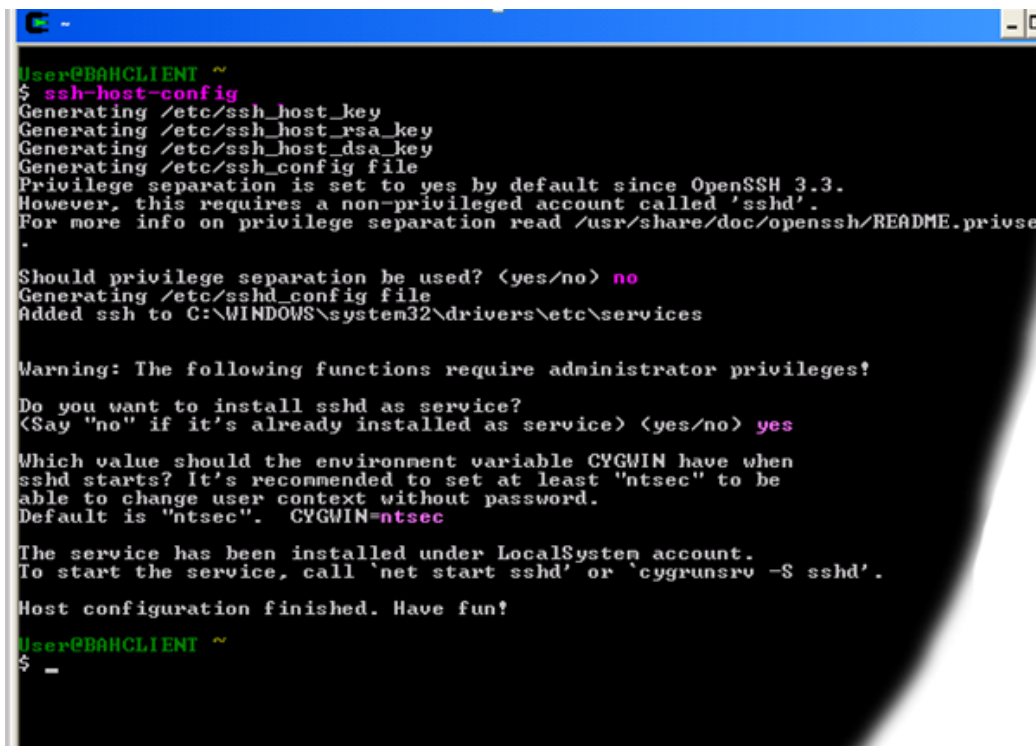
FIGURE A.4 – Modification de la variable path (suite)

A.1.3 Installer le service SSH

Hadoop a besoin des services ssh, cette étape montre comment les intégrer dans Cygwin.

Configuration de ssh daemon

1. Ouvrir le prompt de Cygwin ;
2. Executer les commandes suivantes : `ssh-host-config`
3. privilege separation should be used, répondre `no` .
4. sshd should be installed as a service, répondre `yes` .
5. value of CYGWIN environment variable, tapez `ntsec` , figure A.5.



```

User@BAHCLIENT ~
$ ssh-host-config
Generating /etc/ssh_host_key
Generating /etc/ssh_host_rsa_key
Generating /etc/ssh_host_dsa_key
Generating /etc/ssh_config file
Privilege separation is set to yes by default since OpenSSH 3.3.
However, this requires a non-privileged account called 'sshd'.
For more info on privilege separation read /usr/share/doc/openssh/README.privse
.
Should privilege separation be used? <yes/no> no
Generating /etc/sshd_config file
Added ssh to C:\WINDOWS\system32\drivers\etc\services

Warning: The following functions require administrator privileges!
Do you want to install sshd as service?
<Say "no" if it's already installed as service> <yes/no> yes
Which value should the environment variable CYGWIN have when
sshd starts? It's recommended to set at least "ntsec" to be
able to change user context without password.
Default is "ntsec". CYGWIN=ntsec

The service has been installed under LocalSystem account.
To start the service, call 'net start sshd' or 'cygrunsrv -S sshd'.

Host configuration finished. Have fun!
User@BAHCLIENT ~
$ -

```

FIGURE A.5 – Configuration du service ssh

Démarrer SSH daemon

1. Cliquer sur l'icône `ordinateur` puis choisir l'option `Gérer` dans le menu.
2. Ouvrir `Services and Applications` puis cliquer sur l'option `Services` .
3. Cliquer sur le service `CYGWIN sshd` .

4. Cliquer sur le bouton **Démarrer** , figure A.6
5. Une fenêtre est affichée pour montrer la progression de démarrage du service puis la fenêtre disparaît et l'état de CYGWIN sshd change.

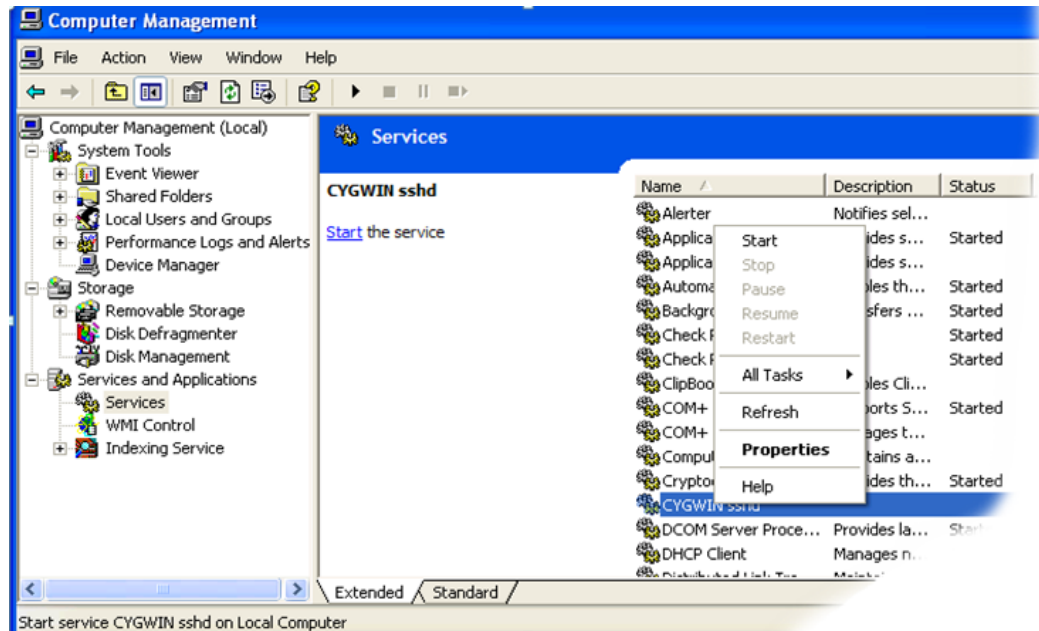


FIGURE A.6 – Démarrer le service ssh

Configuration de la clé d'autorisation

Hadoop nécessite l'authentification de ssh, cette dernière est effectuée via des clés d'autorisation plutôt que des mots de passe. Les étapes suivantes décrivent comment les clés d'autorisation sont configurées, figure A.7

- Ouvrir le prompt de **cygwin**
- Exécuter les commandes suivantes pour générer les clés : **ssh-keygen**
- Puis appuyer sur **Entrée** .
- Une fois la commande est terminée de générer les clés, tapez cette commande pour changer dans le répertoire **ssh** .

```
cd ~\ssh
```

- Tester si les clés ont été générées en exécutants la commande qui permet de lister le contenu du répertoire : **ls -l** Vous devriez trouver les deux fichiers **id_rsa.pub** et **id_rsa** avec une récente création. Ces fichiers contiennent les clés d'autorisation.

```

$ ssh-keygen.exe
Generating public/private rsa key pair.
Enter file in which to save the key (/home/User/.ssh/id_rsa):
Created directory '/home/User/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/User/.ssh/id_rsa.
Your public key has been saved in /home/User/.ssh/id_rsa.pub.
The key fingerprint is:
df:0b:34:58:e3:81:44:c6:e8:e0:af:ea:be:21:a8:5b User@BAHCLIENT

User@BAHCLIENT ~
$ cd .ssh

User@BAHCLIENT ~/.ssh
$ ls -l
total 5
-rw----- 1 User None 1675 Mar 10 09:09 id_rsa
-rw-r--r-- 1 User None 396 Mar 10 09:09 id_rsa.pub

User@BAHCLIENT ~/.ssh
$ cat id_rsa.pub >> authorized_keys

User@BAHCLIENT ~/.ssh
$

```

FIGURE A.7 – Configuration de la clé d'autorization

- Pour enregistrer les nouvelles clés d'autorisation, entrez la commande suivante (notez que les doubles crochets fortement inclinés ils sont très importants) : `cat id_rsa.pub >> authorized_keys`

- Testez si les clés ont été correctement configurées en exécutant cette commande : `ssh localhost`

Comme il s'agit d'une nouvelle installation ssh, vous serez averti que l'authenticité de l'hôte n'a pas pu être établie et on vous demandera si vous voulez vraiment vous connecter.

Répondez oui et appuyez sur **ENTREE** . Vous devriez voir à nouveau l'invite **Cygwin**, ce qui signifie que vous avez réussi à vous connecter.

- Exécutez de nouveau la commande : `ssh localhost`

A.2 TPN°2 : INSTALLATION DE HADOOP

1. Télécharger [Hadoop-2.7.1](#) (ou une autre version supérieure) et décompresser le sur le disque `C:\Hadoop-2.7.1` ;
2. Ouvrir le prompt Cygwin ;
3. Executer les commandes suivantes : `cd` Pour sortir du repertoire `C:\cygwin\.\ssh` ;
4. Executer la commande suivante pour activer le dossier [Home](#) et l'afficher à travers la fenêtre Windows : `explorer .` ;
5. Ouvrir une autre fenêtre et chercher le dossier qui contient le fichier archive Hadoop téléchargé ;
6. Copier le fichier archive de Hadoop dans votre dossier Home, figure [A.8](#).

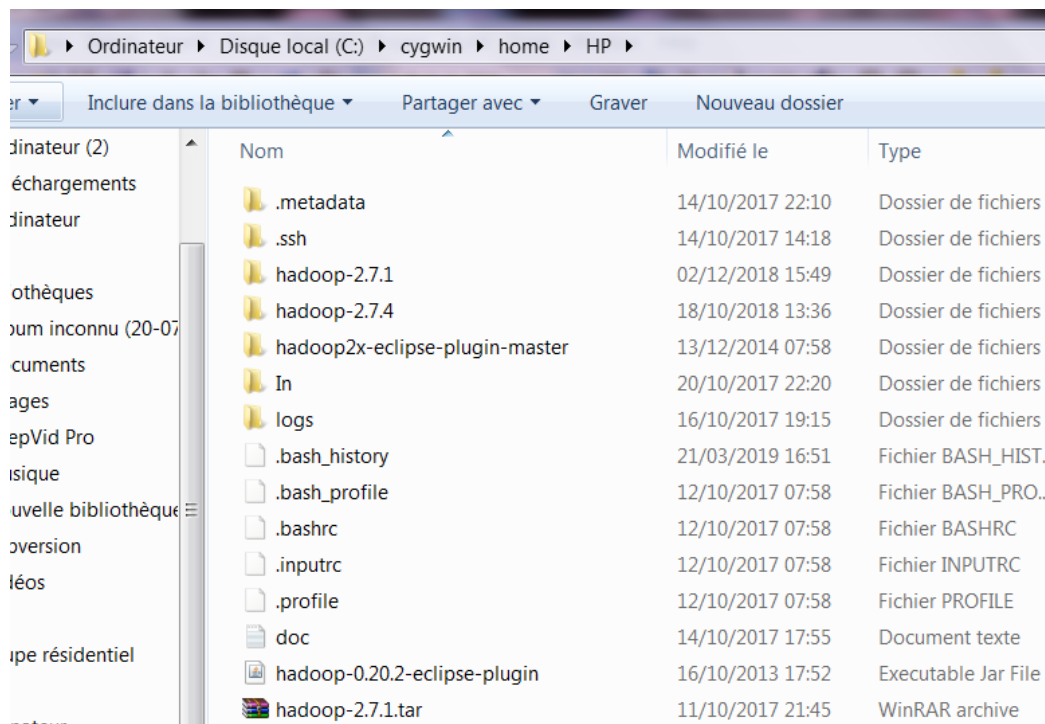
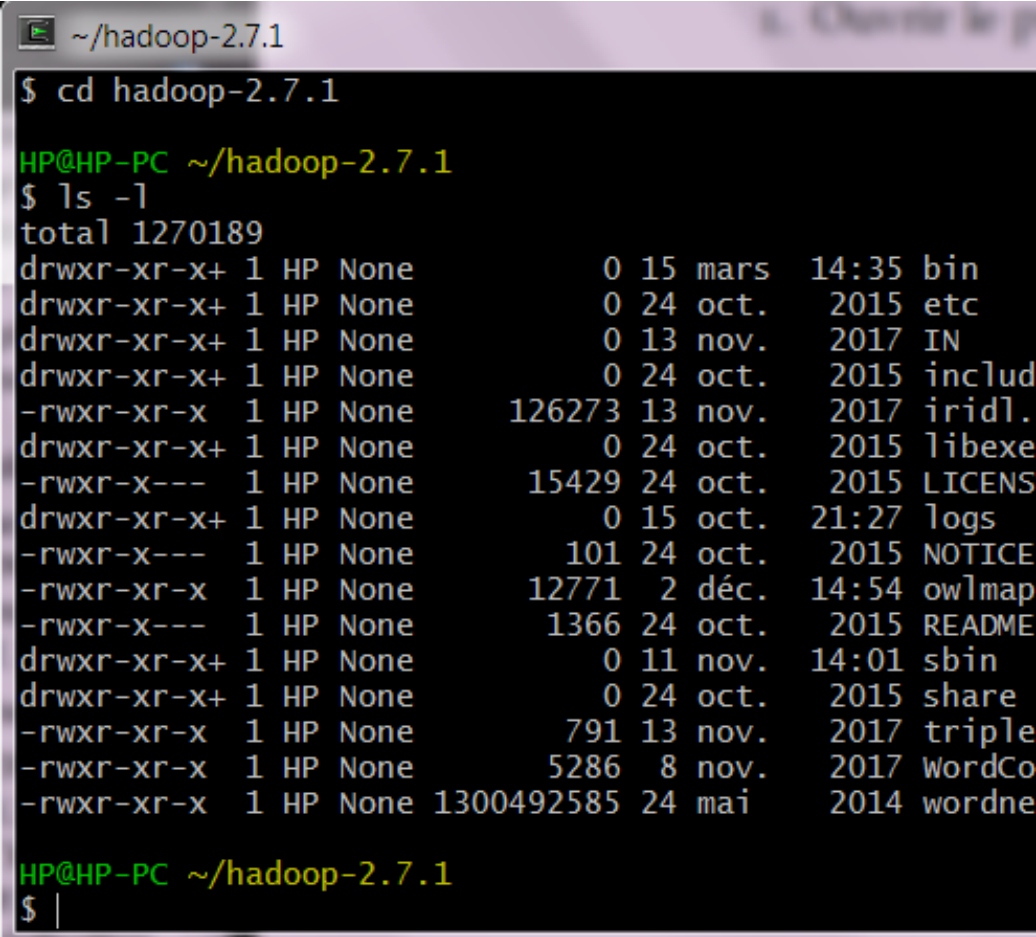


FIGURE A.8 – Copier le fichier archive Hadoop

Pour décompresser le fichier archive suivre les étapes suivantes :

1. Ouvrir le prompt Cygwin ;

2. Tapez la commande suivante : `tar -xzf hadoop-2.7.1.tar.gz` ;
3. L'opération prend quelques minutes, puis vous aurez le prompt Cygwin qui apparaît de nouveau ;
4. Tapez la commande suivante pour afficher le contenu du dossier Home, vous devrez avoir un nouveau dossier intitulé `hadoop-2.7.1` `ls -l` ;
5. Exécutez la commande suivante pour se placer dans le dossier `hadoop-2.7.1` : `cd hadoop-2.7.1` ;
6. Tapez la commande suivante pour afficher le contenu du répertoire `ls -l`, figure A.9.



```

~/hadoop-2.7.1
$ cd hadoop-2.7.1
HP@HP-PC ~/hadoop-2.7.1
$ ls -l
total 1270189
drwxr-xr-x+ 1 HP None          0 15 mars  14:35 bin
drwxr-xr-x+ 1 HP None          0 24 oct.   2015 etc
drwxr-xr-x+ 1 HP None          0 13 nov.   2017 IN
drwxr-xr-x+ 1 HP None          0 24 oct.   2015 includ
-rwxr-xr-x  1 HP None 126273 13 nov.   2017 iridl.
drwxr-xr-x+ 1 HP None          0 24 oct.   2015 libexe
-rwxr-x---  1 HP None 15429 24 oct.   2015 LICENS
drwxr-xr-x+ 1 HP None          0 15 oct.  21:27 logs
-rwxr-x---  1 HP None 101 24 oct.   2015 NOTICE
-rwxr-xr-x  1 HP None 12771  2 déc.   14:54 owlmap
-rwxr-x---  1 HP None 1366 24 oct.   2015 README
drwxr-xr-x+ 1 HP None          0 11 nov.   14:01 sbin
drwxr-xr-x+ 1 HP None          0 24 oct.   2015 share
-rwxr-xr-x  1 HP None 791 13 nov.   2017 triple
-rwxr-xr-x  1 HP None 5286  8 nov.   2017 WordCo
-rwxr-xr-x  1 HP None 1300492585 24 mai   2014 wordne
HP@HP-PC ~/hadoop-2.7.1
$ |

```

FIGURE A.9 – Copier le fichier archive Hadoop

Compléter l'installation avec la commande suivante :

ajouter le chemin `C : \cygwin\home\HP\hadoop-2.7.4\bin` et `C : \cygwin\home\HP\hadoop-2.7.4\sbin` à la variable d'environnement `path`.

Configuration et lancement de Hadoop

Il faut maintenant définir la configuration de Hadoop et pour cela plusieurs fichiers de configuration doivent être modifiés. Dans Hadoop, les fichiers de configuration fonctionnent sur le principe de (clé,valeur) : la clé correspondant au nom du paramètre et la valeur est celle assignée à ce paramètre, tout cela au format XML.

1. Il faut tout d'abord configurer Hadoop en mode nœud unique (local) en éditant le fichier `etc\hadoop\core-site.xml` de la manière suivante, figure A.10 ;
2. Le fichier `etc\hadoop\hdfs-site.xml` contient les paramètres spécifiques au système de fichiers HDFS, (figure A.11), avec le nombre de répllication d'un bloc (qui vaut 1 ici) ;
3. Il faut ensuite configurer les paramètres spécifiques à MapReduce qui sont dans le fichier `etc\hadoop\mapred-site.xml` , figure A.12. Ici, on précise que YARN est utilisé comme implémentation de MapReduce ;
4. Hadoop est désormais correctement installé et configuré. Il reste juste à formater le système de fichiers HDFS local, figure A.13 : `hdfs namenode -format`
5. et à démarrer Hadoop avec :
 - `start-dfs.sh`
 - `start-yarn.sh`

```
18
19 <configuration>
20   <property>
21     <name>fs.defaultFS</name>
22     <value>hdfs://localhost:9000</value>
23   </property>
24 </configuration>
```

FIGURE A.10 – le fichier *etc\hadoop\core-site.xml*

```
18
19 <configuration>
20   <property>
21     <name>dfs.replication</name>
22     <value>1</value>
23   </property>
24 </configuration>
```

FIGURE A.11 – le fichier *etc\hadoop\hdfs-site.xml*

```
18
19 <configuration>
20   <property>
21     <name>mapreduce.framework.name</name>
22     <value>yarn</value>
23   </property>
24 </configuration>
```

FIGURE A.12 – le fichier *etc\hadoop\mapred-site.xml*

```
HP@HP-PC ~/hadoop-2.7.1
$ hdfs namenode -format
19/03/21 18:21:03 INFO namenode.NameNode: STARTUP_MSG:
/*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG:  host = HP-PC/192.168.1.5
STARTUP_MSG:  args = [-format]
STARTUP_MSG:  version = 2.7.1
STARTUP_MSG:  classpath = C:\hadoop-2.7.1\etc\hadoop;C:\hadoop-2.7.1\share\hado
op\common\lib\activation-1.1.jar;C:\hadoop-2.7.1\share\hadoop\common\lib\apached
s-i18n-2.0.0-M15.jar;C:\hadoop-2.7.1\share\hadoop\common\lib\apacheds-kerberos-c
odec-2.0.0-M15.jar;C:\hadoop-2.7.1\share\hadoop\common\lib\api-asn1-api-1.0.0-M2
0.jar;C:\hadoop-2.7.1\share\hadoop\common\lib\api-util-1.0.0-M20.jar;C:\hadoop-2
.7.1\share\hadoop\common\lib\asm-3.2.jar;C:\hadoop-2.7.1\share\hadoop\common\lib
\avro-1.7.4.jar;C:\hadoop-2.7.1\share\hadoop\common\lib\commons-beanutils-1.7.0.
jar;C:\hadoop-2.7.1\share\hadoop\common\lib\commons-beanutils-core-1.8.0.jar;C:\
hadoop-2.7.1\share\hadoop\common\lib\commons-cli-1.2.jar;C:\hadoop-2.7.1\share\h
adoop\common\lib\commons-codec-1.4.jar;C:\hadoop-2.7.1\share\hadoop\common\lib\c
ommons-collections-3.2.1.jar;C:\hadoop-2.7.1\share\hadoop\common\lib\commons-com
press-1.4.1.jar;C:\hadoop-2.7.1\share\hadoop\common\lib\commons-configuration-1.
6.jar;C:\hadoop-2.7.1\share\hadoop\common\lib\commons-digester-1.8.jar;C:\hadoop
-2.7.1\share\hadoop\common\lib\commons-httpclient-3.1.jar;C:\hadoop-2.7.1\share\
hadoop\common\lib\commons-io-2.4.jar;C:\hadoop-2.7.1\share\hadoop\common\lib\com
```

FIGURE A.13 – *Formater HDFS*

BIBLIOGRAPHIE

- [18] vlad korolev. hadoop on windows with eclipse. 2008. URL <http://v-lad.org/Tutorials/Hadoop/>.