

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^{eme} année

Dr. Klai Sihem

AVANT PROPOS...

Le Big Data est une science récente qui a surgie avec l'évolution et la variation des données et d'applications mises et échangées en ligne. Cette science consiste à prendre en charge d'une manière efficace un volume important des données hétérogènes en intégrant des techniques et outils nouveaux, vu que la technologie disponible ne répond plus aux besoins.

Ce polycopié est un support pédagogique qui permet d'initier l'étudiant au domaine des Big Data. Ce cours composé de plusieurs chapitres permet aux étudiants de comprendre la problématique et la motivation du domaine, et de maîtriser l'outil Hadoop avec le modèle MapReduce associé à ce domaine.

Chaque chapitre est élaboré pour répondre à un but pédagogique bien précis, se matérialisant par des explications, définitions accompagnées d'exemples et des illustrations par des figures suivies par des exercices, des solutions envisageables ou des fiches de travaux pratiques bien guidés.

1. Le chapitre I met l'étudiant dans le contexte du Big Data, consiste à lui donner des connaissances générales sur le domaine ;
2. Le chapitre II est consacré à l'étude de Hadoop, le framework qui permet le développement d'applications traitant les données massives. Ce chapitre donne les notions les plus générales avec la procédure d'installation du logiciel ;
3. Le chapitre III détaille la partie qui s'occupe du stockage des données "HDFS", avec la possibilité de la manipulation de ces données selon deux manières différentes à savoir : les commandes et l'API JAVA ;

4. Le chapitre IV étudie en détail la partie traitement des données massives "MapReduce", le modèle qui permet de traiter des blocs de données séparément et parallèlement dans des machines connectées. La modélisation selon le paradigme MapReduce est une étape importante avant le développement des programmes ;
5. Le chapitre V détaille l'implémentation des programmes MapReduce dans Hadoop. Dans le cadre de ce chapitre, nous étudions l'implémentation des programmes en utilisant le langage Java. D'autres langages peuvent être utilisés pour écrire des programmes mapreduce, mais cette partie n'est pas traitée dans ce cours.

L'élaboration de ce polycopié a été inspirée de plusieurs documents, j'ai pris le soin de les citer dans la partie bibliographie, d'autres documents aussi ont été cité afin d'apporter aux lecteurs plus de commandes et plus de détails sur les parties traitées dans ce cours.

TERMINOLOGIE . . .

JVM : Java virtuelle machine

HDFS : Hadoop Distributed File System

YARN : Yet Another Ressource Negotiator

AM : Application Master

API java : Application Programming Interfaces java

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 . . .	16
BIBLIOGRAPHIE	21

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 Démarer le service ssh	6
A.7 Configuration de la clé d'authorization	7
A.8 Copier le fichier archive Hadoop	8
A.9 Copier le fichier archive Hadoop	9
A.10le fichier <code>etc\hadoop\core-site.xml</code>	11
A.11le fichier <code>etc\hadoop\hdfs-site.xml</code>	11
A.12le fichier <code>etc\hadoop\mapred-site.xml</code>	11
A.13Formater HDFS	12
A.14Commande hdfs	13
A.15Commande hdfs dfs	13
A.16Commande hdfs dfs -cat appel-T.txt	14
A.17Commande hdfs dfs -put *.txt entree	14
A.18Commande hdfs dfs -rm in.txt	15
A.19Commande hdfs dfs -mkdir entree	15
A.20Insertion des fichiers Jar Hadoop dans le projet	17
A.21Insertion du programme de l'exercie1 du chapitre ?? dans Eclipse	17
A.22Génération du fichier JAR	18
A.23Vérification si le programme .jar généré existe	18
A.24Exécution du programme	19

A.25 Affichage des résultats	19
--	----

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 [18](#) :

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.

A. ANNEXES



FIGURE A.1 – Installation de Cygwin

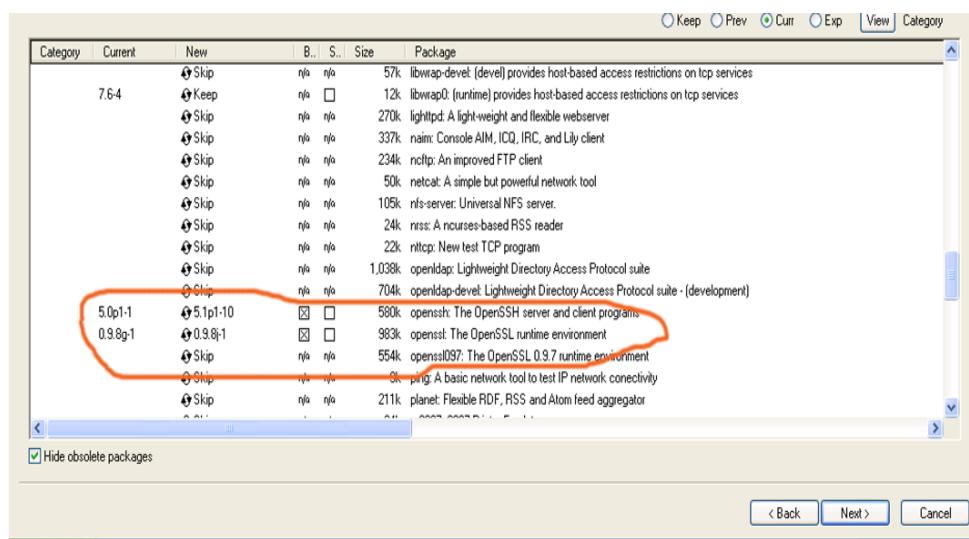


FIGURE A.2 – Installation de openssh

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 boite de dialogue A.3 apparait 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 boites de dialogues.

A. ANNEXES

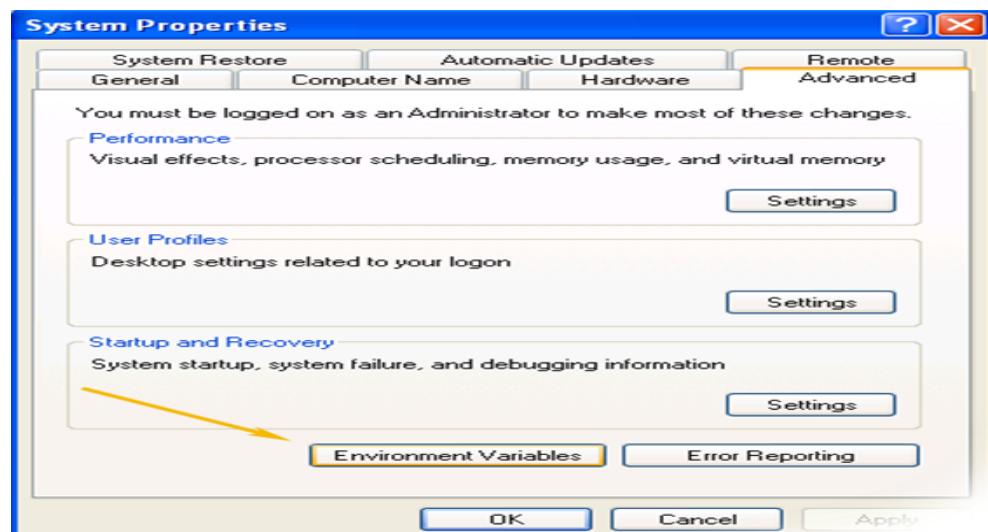


FIGURE A.3 – Modification de la variable path

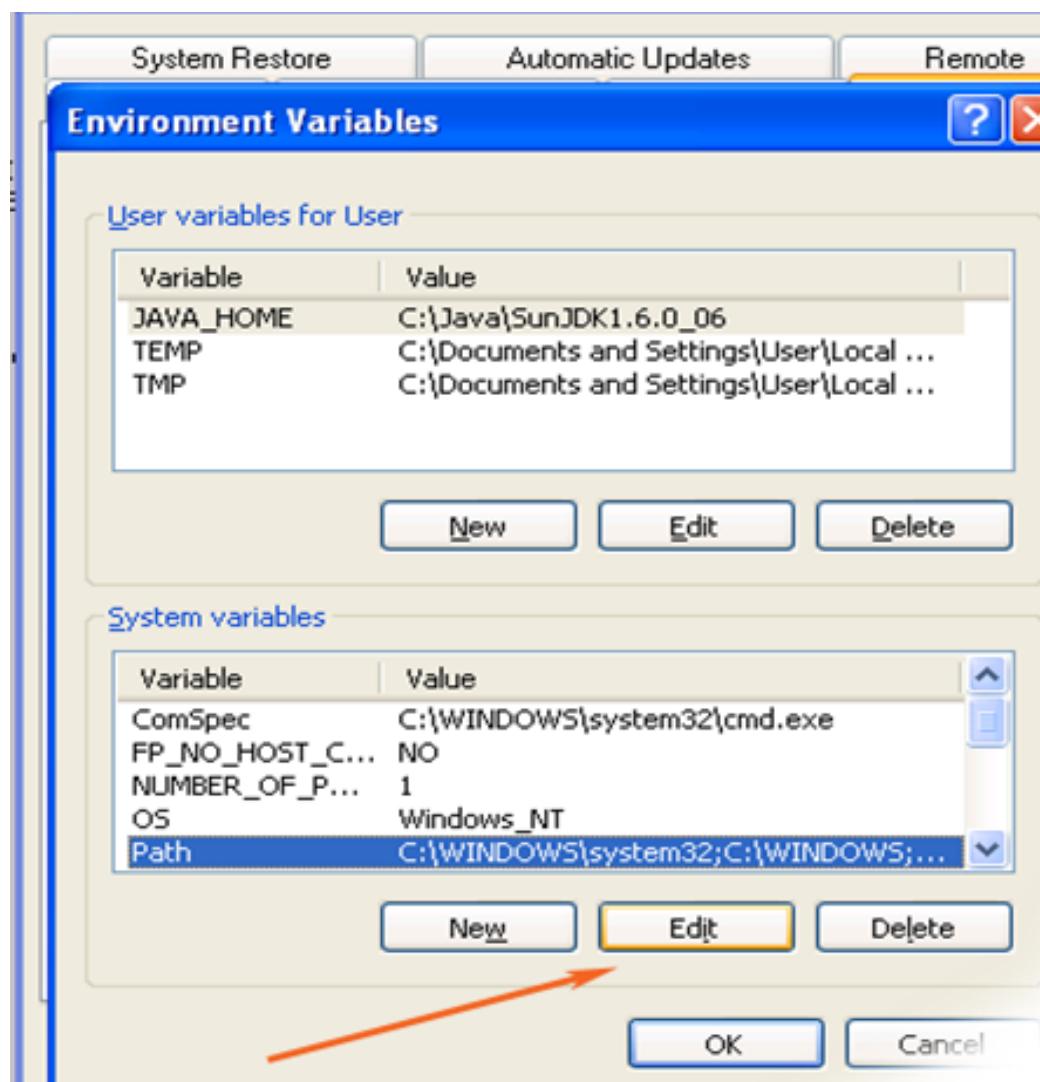


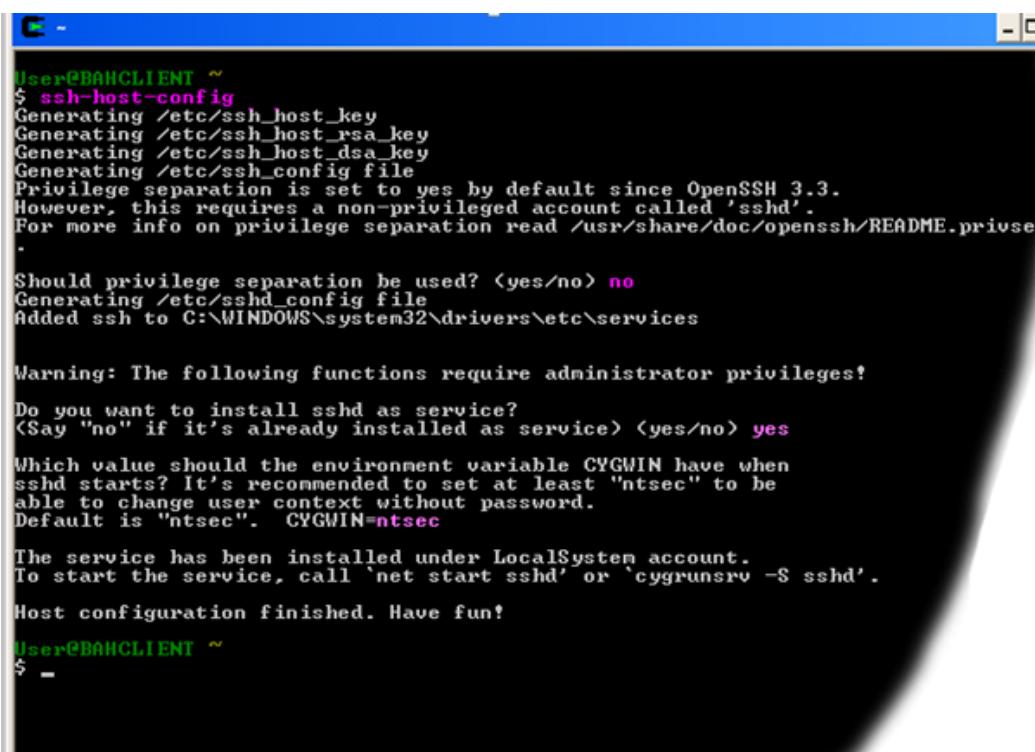
FIGURE A.4 – Modification de la variable path (suite)

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.privsep

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.

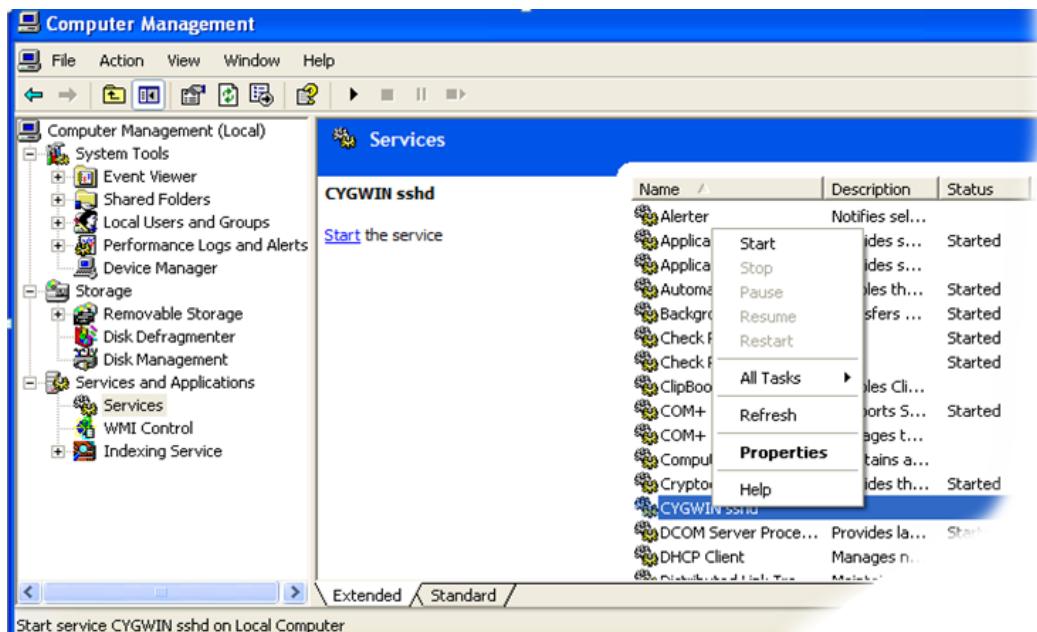


FIGURE A.6 – Démarrer le service ssh

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 devrez 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: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'authorization

- 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.

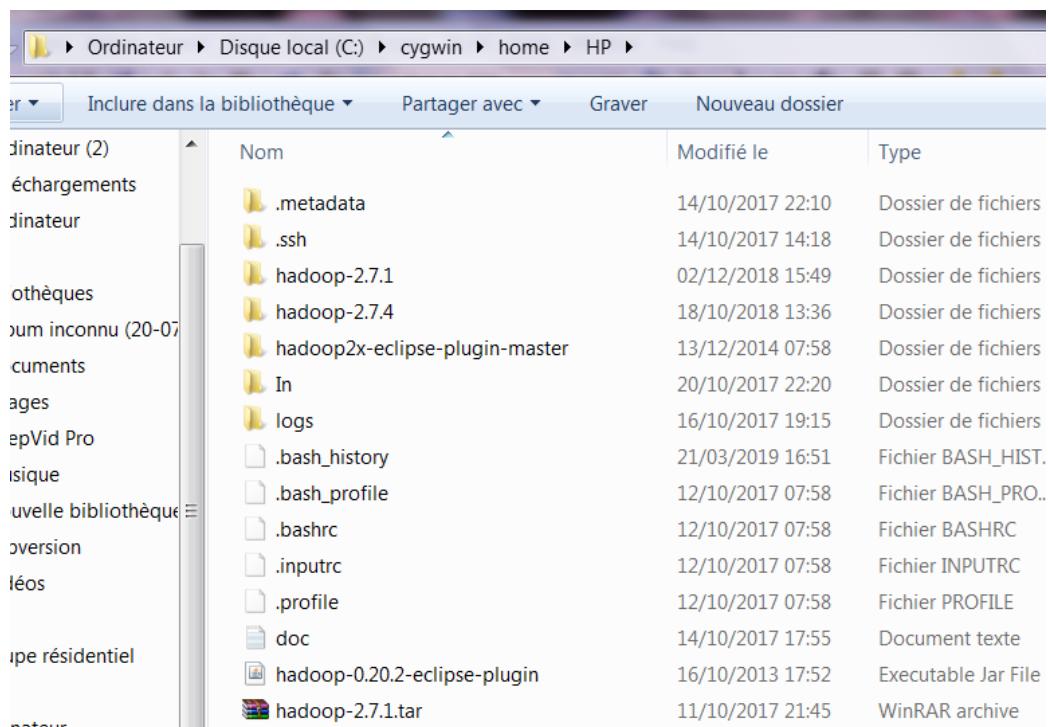
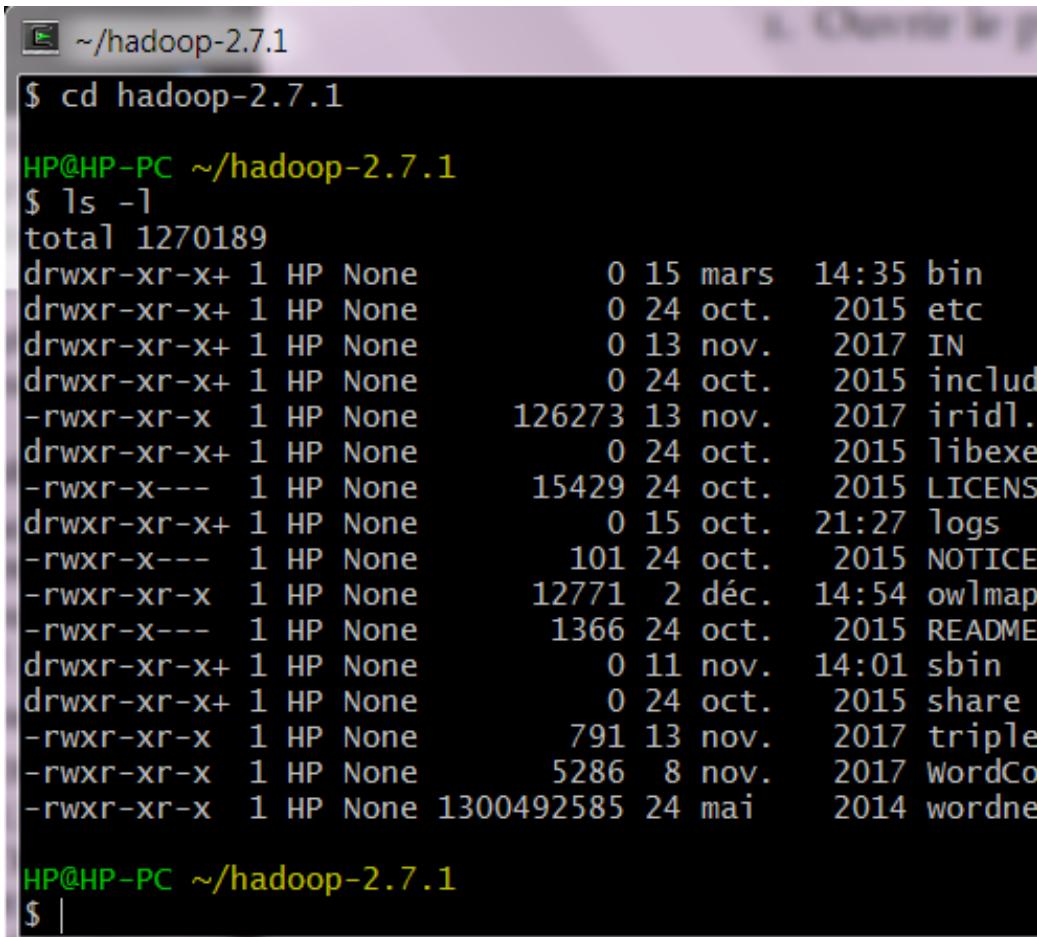


FIGURE A.8 – Copier le fichier archive Hadoop

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 apparait 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. Executez 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
-rw xr-xr-x 1 HP None         126273 13 nov.  2017 iridl.
drwxr-xr-x+ 1 HP None          0 24 oct.   2015 libexe
-rw xr-x--- 1 HP None         15429  24 oct.  2015 LICENS
drwxr-xr-x+ 1 HP None          0 15 oct.  21:27 logs
-rw xr-x--- 1 HP None          101  24 oct.  2015 NOTICE
-rw xr-xr-x 1 HP None         12771  2 déc.  14:54 owlmap
-rw xr-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
-rw xr-xr-x 1 HP None          791  13 nov.  2017 triple
-rw xr-xr-x 1 HP None          5286  8 nov.  2017 WordCo
-rw xr-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 réPLICATION 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\hadoop\common\lib\activation-1.1.jar;C:\hadoop-2.7.1\share\hadoop\common\lib\apacheds-i18n-2.0.0-M15.jar;C:\hadoop-2.7.1\share\hadoop\common\lib\apacheds-kerberos-codect-2.0.0-M15.jar;C:\hadoop-2.7.1\share\hadoop\common\lib\api-asn1-api-1.0.0-M20.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\hadoop\common\lib\commons-collections-3.2.1.jar;C:\hadoop-2.7.1\share\hadoop\common\lib\commons-compress-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

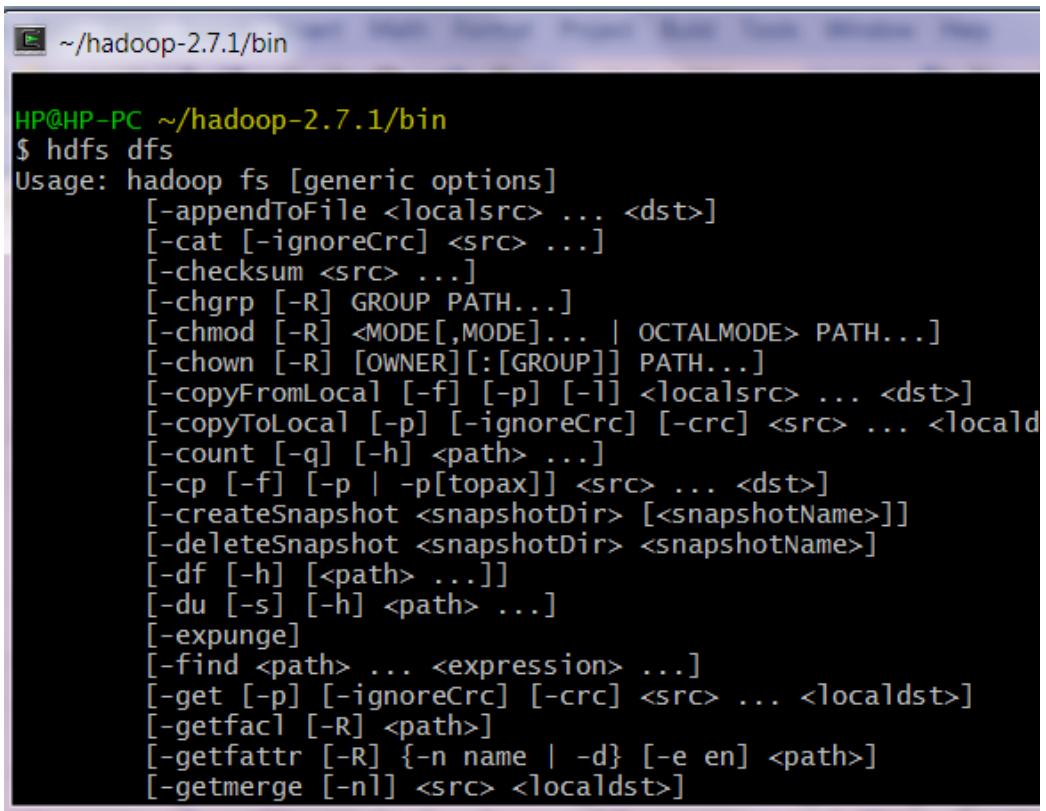
A.3 TPN°3 : SE FAMILIARISER AVEC LES COMMANDES HDFS

Ouvrez le prompt Cygwin et tapez les commandes suivantes :

- **hdfs** : permet d'afficher les commandes associées à hdfs, figure A.14
- **hdfs dfs** : permet d'afficher les commandes du nouveau système de fichier mis à notre disposition, figure A.15
- **hdfs dfs -cat appel-T.txt** : permet d'afficher le contenu du fichier "appel-T".txt, figure A.16
- **hdfs dfs -put *.txt entree** : permet de copier les fichiers ".txt" dans le répertoire "entree" A.17
- **hdfs dfs -rm in.txt** : permet de supprimer le fichier "in.txt" A.18
- **hdfs dfs -mkdir entree** : permet de créer le répertoire "entree", puis tapez ls pour vérifier que le dossier "entree" est créé, figure A.19
- d'autres commandes seront testées pendant le TP.

```
HP@HP-PC ~/hadoop-2.7.1/bin
$ hdfs
Usage: hdfs [--config confdir] [--loglevel loglevel] COMMAND
        where COMMAND is one of:
          dfs                         run a filesystem command on the file systems su
Hadoop.
          classpath                   prints the classpath
          namenode -format            format the DFS filesystem
          secondarynamenode           run the DFS secondary namenode
          namenode                    run the DFS namenode
          journalnode                run the DFS journalnode
          zkfc                        run the ZK Failover Controller daemon
          datanode                   run a DFS datanode
          dfsadmin                   run a DFS admin client
          haadmin                     run a DFS HA admin client
          fsck                        run a DFS filesystem checking utility
          balancer                   run a cluster balancing utility
          jmxget                      get JMX exported values from NameNode or DataNo
          mover                       run a utility to move block replicas across
                                      storage types
          oiv                         apply the offline fsimage viewer to an fsimage
          oiv_legacy                  apply the offline fsimage viewer to an legacy f
          oev                         apply the offline edits viewer to an edits file
```

FIGURE A.14 – Commande *hdfs*



```
~/hadoop-2.7.1/bin
HP@HP-PC ~/hadoop-2.7.1/bin
$ hdfs dfs
Usage: hadoop fs [generic options]
      [-appendToFile <localsrc> ... <dst>]
      [-cat [-ignoreCrc] <src> ...]
      [-checksum <src> ...]
      [-chgrp [-R] GROUP PATH...]
      [-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]
      [-chown [-R] [OWNER][:[GROUP]] PATH...]
      [-copyFromLocal [-f] [-p] [-l] <localsrc> ... <dst>]
      [-copyToLocal [-p] [-ignoreCrc] [-crc] <src> ... <locald
      [-count [-q] [-h] <path> ...]
      [-cp [-f] [-p | -p[topax]] <src> ... <dst>]
      [-createSnapshot <snapshotDir> [<snapshotName>]]
      [-deleteSnapshot <snapshotDir> <snapshotName>]
      [-df [-h] [<path> ...]]
      [-du [-s] [-h] <path> ...]
      [-expunge]
      [-find <path> ... <expression> ...]
      [-get [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
      [-getfacl [-R] <path>]
      [-getattr [-R] {-n name | -d} [-e en] <path>]
      [-getmerge [-n1] <src> <localdst>]
```

FIGURE A.15 – Commande *hdfs dfs*

```
HP@HP-PC ~/hadoop-2.7.1/bin/entree
$ hdfs dfs -cat appel-T.txt
id1,07782345,01/01/2018,5
id1,06785645,04/05/2018,10
id2,06785645,04/05/2018,10
id1,05678712,01/10/2018,10
id2,06764523,11/06/2018,16
id1,05345645,12/02/2018,23
HP@HP-PC ~/hadoop-2.7.1/bin/entree
$ |
```

FIGURE A.16 – Commande *hdfs dfs -cat appel-T.txt*

```
HP@HP-PC ~/hadoop-2.7.1/bin/entree
$ cd ..
HP@HP-PC ~/hadoop-2.7.1/bin
$ hdfs dfs -put *.txt entree

HP@HP-PC ~/hadoop-2.7.1/bin
$ cd entree

HP@HP-PC ~/hadoop-2.7.1/bin/entree
$ ls
appel-T.txt  input.txt  triple.txt  youtubedata.txt
in.txt       Sortie.txt  uber.txt

HP@HP-PC ~/hadoop-2.7.1/bin/entree
$ |
```

FIGURE A.17 – Commande *hdfs dfs -put *.txt entree*

```
HP@HP-PC ~/hadoop-2.7.1/bin/entree
$ hdfs dfs -rm in.txt
19/03/22 09:15:44 INFO Configuration.deprecated. Instead, use dfs.bytes-per-checksum
19/03/22 09:15:44 INFO fs.TrashPolicyDefault
tion interval = 0 minutes, Emptier interval =
Deleted in.txt

HP@HP-PC ~/hadoop-2.7.1/bin/entree
$ ls
appel-T.txt  input.txt  Sortie.txt  triple.txt

HP@HP-PC ~/hadoop-2.7.1/bin/entree
$ |
```

\item \color{blue}{literated} \textcolor{blue}{\{ hdfs dfs -rm in\\$}

FIGURE A.18 – Commande *hdfs dfs -rm in.txt*

```
HP@HP-PC ~/hadoop-2.7.1/bin
$ hdfs dfs -mkdir entree

HP@HP-PC ~/hadoop-2.7.1/bin
$ cd entree

HP@HP-PC ~/hadoop-2.7.1/bin/entree
$ cd ..
```

FIGURE A.19 – Commande *hdfs dfs -mkdir entree*

A.4 TPN° 4 : IMPLÉMENTER UN PROGRAMME MAPREDUCE : EXERCICE1

On va implémenter le programme Mapreduce de l'exercice 1 du chapitre?? sous Eclipse. Pour cela, suivre les étapes suivantes :

1. Lancer Eclipse;
2. Créer un projet "nom projet" exemple : "A-bon-tel";
3. Importer les packages Hadoop avec **Build Path / Configure Build Path / Add External Jars** puis sélectionner les fichiers Jar de Hadoop, figure A.20;
4. Créer une classe main "nom classe", exemple : "AppelTelp" c'est aussi la classe Driver;
5. Ecrire le programme de l'exercice1 du chapitre ?? dans la classe, figure A.21;
6. Sélectionner le **File**, cliquez sur **Export**, puis **Jar file**, sélectionnez votre projet; insérez le chemin de sortie avec le nom du fichier Jar à générer, puis cliquez sur **Finish** pour terminer A.22;
7. Ouvrez le prompt Cygwin, se déplacer vers le sous répertoire bin de hadoop;
8. Vérifier si le fichier jar généré existe bien, figure A.23;
9. Insérer le fichier en entrée et afficher le contenu pour vérification, figure A.16
10. Exécutez le programme avec la commande :
`hadoop jar abontel.jar AppelTelp appel-T.txt outappelT`, figure A.24
 - **abontel.jar** est le fichier jar qu'on vient de créer;
 - **AppelTelp** est le nom de la classe main ou Driver;
 - **appel-T.txt** est le nom du fichier en entrée;
 - **outappelT** est le nom du répertoire en sortie.
11. affichons maintenant le résultat :
 - Se déplacer dans le répertoire dans "outappelT";

A. ANNEXES

- Tapez ls ls pour afficher le contenu du répertoire de sortie ;
- Le résultat est sauvegardé dans le fichier "part-r-oooooo" ;
- Afficher le contenu de ce fichier et vérifier le résultat, figure A.25.

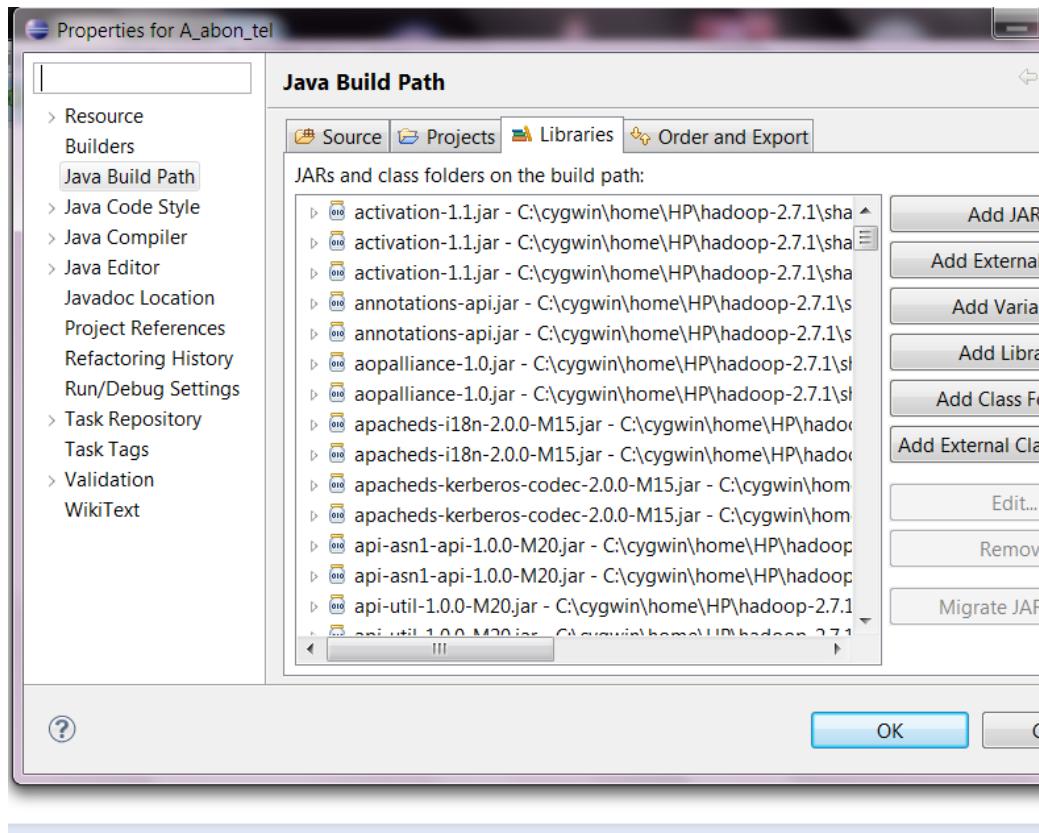


FIGURE A.20 – Insertion des fichiers Jar Hadoop dans le projet

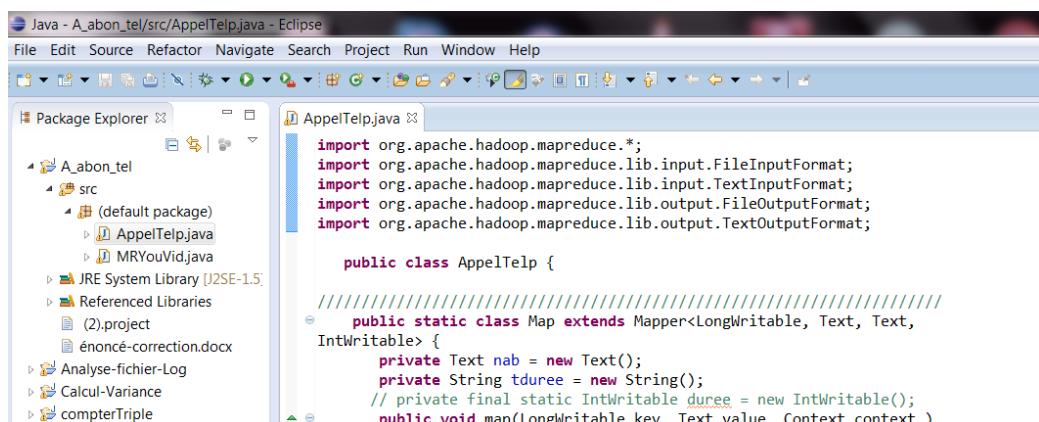


FIGURE A.21 – Insertion du programme de l'exercice1 du chapitre ?? dans Eclipse

A. ANNEXES

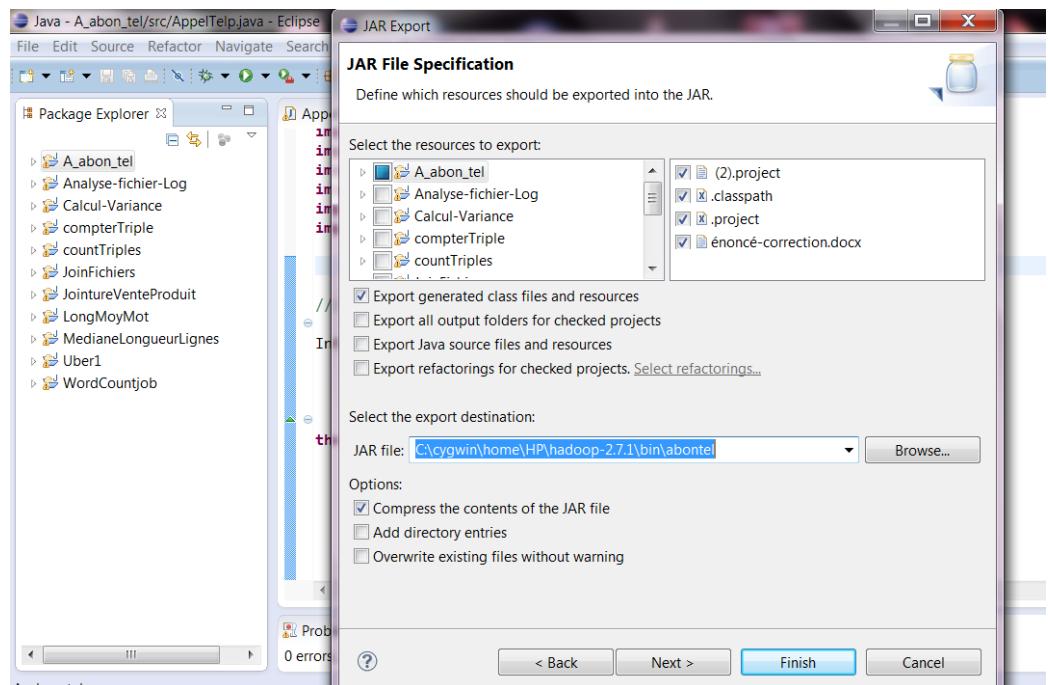
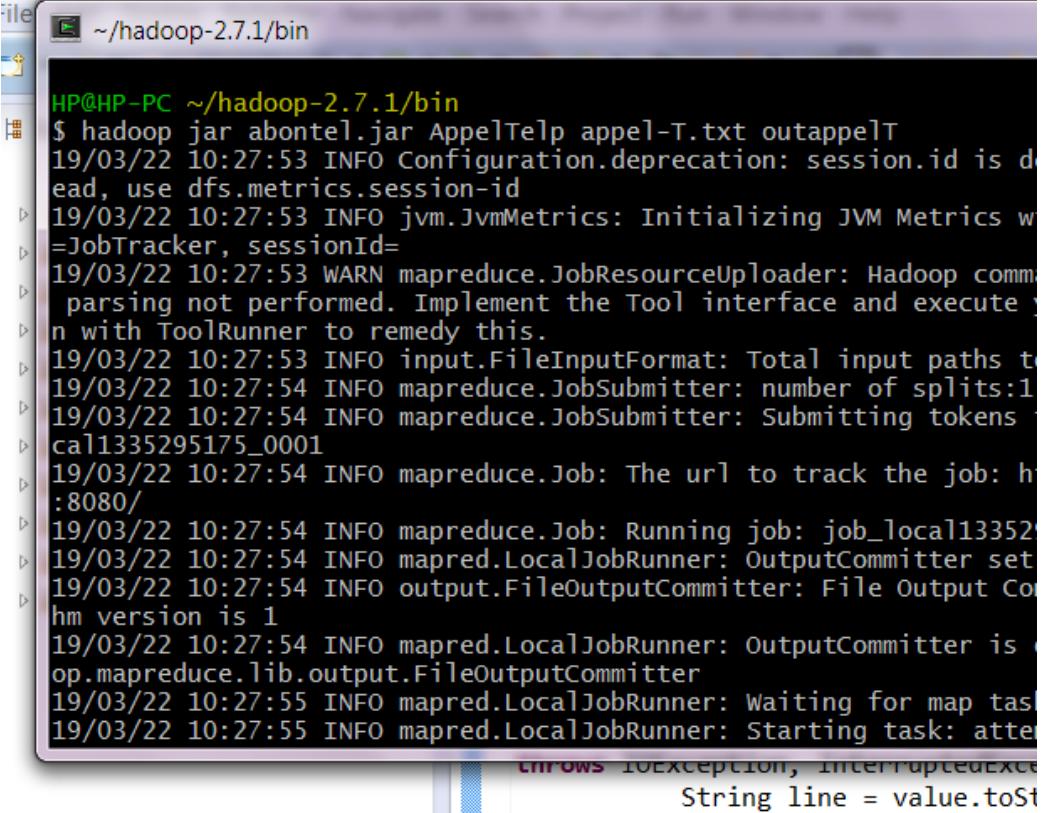


FIGURE A.22 – Génération du fichier JAR

```
HP@HP-PC ~ /hadoop-2.7.1/bin
$ ls *.jar
abontel.jar          LA.jar
abt.jar              LM.jar
AppelTelp.jar        LogAnalyse.j
compter.jar          LogAnalysem.j
CompterTriple.jar    LongM.jar
CountTriples.jar    mono.jar
```

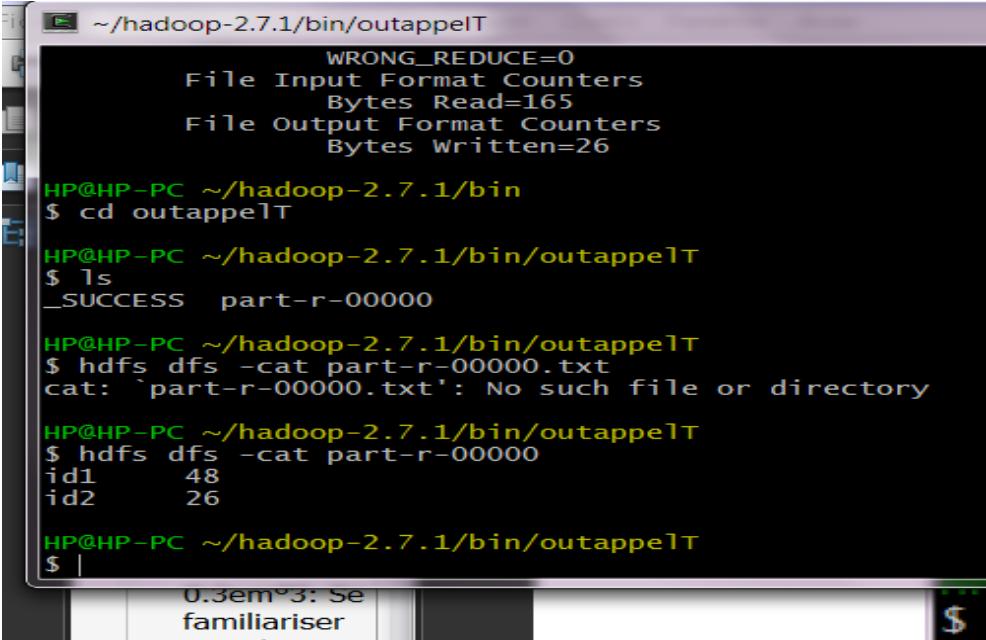
FIGURE A.23 – Vérification si le programme .jar généré existe

A. ANNEXES



```
HP@HP-PC ~/hadoop-2.7.1/bin
$ hadoop jar abontel.jar AppelTelp appel-T.txt outappelt
19/03/22 10:27:53 INFO Configuration.deprecation: session.id is deprecated, use dfs.metrics.session-id
19/03/22 10:27:53 INFO jvm.JvmMetrics: Initializing JVM Metrics with configuration: {}
19/03/22 10:27:53 WARN mapreduce.JobResourceUploader: Hadoop command parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
19/03/22 10:27:53 INFO input.FileInputFormat: Total input paths to process : 1
19/03/22 10:27:54 INFO mapreduce.JobSubmitter: number of splits:1
19/03/22 10:27:54 INFO mapreduce.JobSubmitter: Submitting tokens for application application_1335295175_0001
19/03/22 10:27:54 INFO mapreduce.Job: The url to track the job: http://127.0.0.1:8080/
19/03/22 10:27:54 INFO mapreduce.Job: Running job: job_local1335295175_0001
19/03/22 10:27:54 INFO mapred.LocalJobRunner: OutputCommitter set
19/03/22 10:27:54 INFO output.FileOutputCommitter: File Output Committer version is 1
19/03/22 10:27:54 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
19/03/22 10:27:55 INFO mapred.LocalJobRunner: Waiting for map tasks
19/03/22 10:27:55 INFO mapred.LocalJobRunner: Starting task: attempt_1335295175_0001_m_000000_0
```

FIGURE A.24 – Exécution du programme



```
HP@HP-PC ~/hadoop-2.7.1/bin/outappelt
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=165
File Output Format Counters
Bytes Written=26
HP@HP-PC ~/hadoop-2.7.1/bin
$ cd outappelt
HP@HP-PC ~/hadoop-2.7.1/bin/outappelt
$ ls
_SUCCESS  part-r-00000
HP@HP-PC ~/hadoop-2.7.1/bin/outappelt
$ hdfs dfs -cat part-r-00000.txt
cat: 'part-r-00000.txt': No such file or directory
HP@HP-PC ~/hadoop-2.7.1/bin/outappelt
$ hdfs dfs -cat part-r-00000
id1      48
id2      26
HP@HP-PC ~/hadoop-2.7.1/bin/outappelt
$ |
```

FIGURE A.25 – Affichage des résultats

BIBLIOGRAPHIE

- [01]. site : Le big data, 2019. URL <https://www.lebigdata.fr/definition-big-data>.
- [02]. Stephane Vialle, CentraleSupélec. Big Data : Informatique pour les données et calculs massifs. 24 mai 2017 .
- [03]. mohammed zuhair al taie. hadoop ecosystem : an integrated environment for big data. in big data, guest posts 2015. URL <http://blog.agroknow.com/?p=3810>.
- [04]. site : Hadoop. 2016. URL <http://www.opentuto.com/category/web-2/big-data/hadoop/>.
- [05]. amal abid. cours big data : Chapitre2 : Hadoop. 2017. URL <https://fr.slideshare.net/AmalAbid1/cours-big-data-chap2>.
- [06]. celine hudelot, regis behmo. realisez des calculs distribue sur des donnees massives. 2019. URL <https://openclassrooms.com/fr/courses/4297166-realisez-des-calculs-distribues-sur-des-donnees-massives>.
- [07]. Benaouda, Sid Ahmed Amine, implantation du modèle mapreduce dans l'environnement distribué, 2015 . URL <http://dspace.univ-tlemcen.dz>.
- [08]. Tom White, Hadoop : The Definitive Guide, June 2009 : First Edition.
- [09]. Srinath Perera, Thilina Gunarathne, Hadoop MapReduce Cookbook, First published : February 2013.

- [10]. Pierre Nerzic, Outils Hadoop pour le BigData, mars 2018.
- [11]. marty hall. map reduce 2.0 (input and output). 2013. URL <https://www.slideshare.net/martyhall/hadoop-tutorial-mapreduce-part-4-input-and-output>.
- [12]. build projects, learn skills, get hired. hadoop mapreduce- java-based processing framework for big data. URL <https://www.dezyre.com/hadoop-course/mapreduce>.
- [13]. URL <https://hadoop.apache.org/docs/r2.4.1/api/org/apache/hadoop/mapreduce/>.
- [14]. URL <https://gist.github.com/kzk/712029/9d0833aac03b23ec226e034d98f5871d9580724e>.
- [15]. Jeffrey Dean and Sanjay Ghemawat, MapReduce : Simplified Data Processing on Large Clusters, 2004 .
- [16]. Univ. Lille1- Licence info 3eme annee, Cours n°1 : Introduction à la programmation fonctionnelle, 2013-2014 .
- [17] dataflair team. hadoop architecture in detail hdfs,yarn et mapreduce. 2019. URL <https://data-flair.training/blogs/hadoop-architecture/>.
- [18] vlad korolev. hadoop on windows with eclipse. 2008. URL <http://v-lad.org/Tutorials/Hadoop/>. (Cité page 2.)
- [19]. Donald Miner and Adam Shook. MapReduce Design Patterns. O'Reilly, 2012.
- [20]. Tom White. Hadoop : The Definitive Guide, 4th Edition. O'Reilly, 2015.
- [21]. Hadoop Training. 2018. URL <https://www.slideshare.net/AnandMHadoop/session-19-mapreduce>.