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 traiteé 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 Negociator

AM : Application Master

API java : Application Programming Interfaces java

Table des matières

Liste des figures

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.

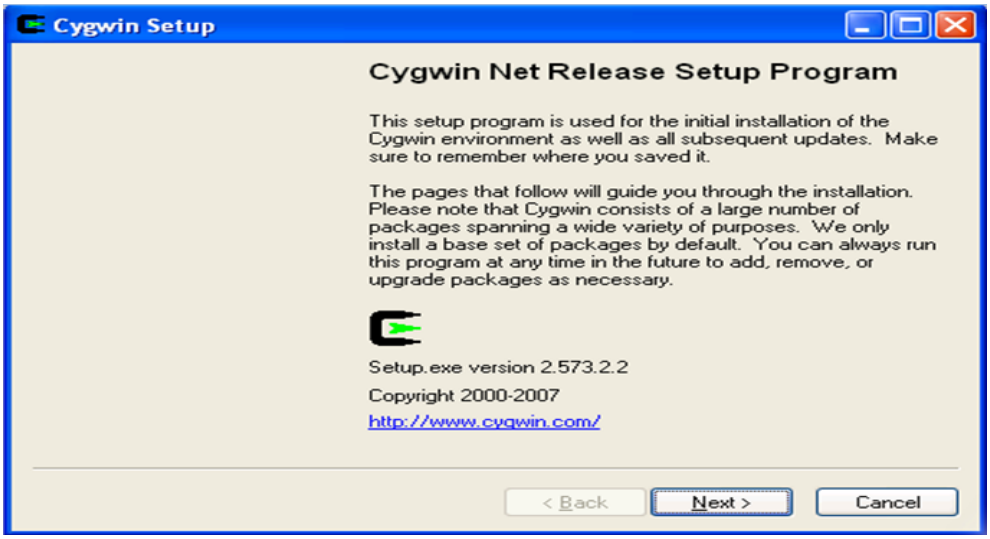

FIGURE A.1 – *Installation de Cygwin*

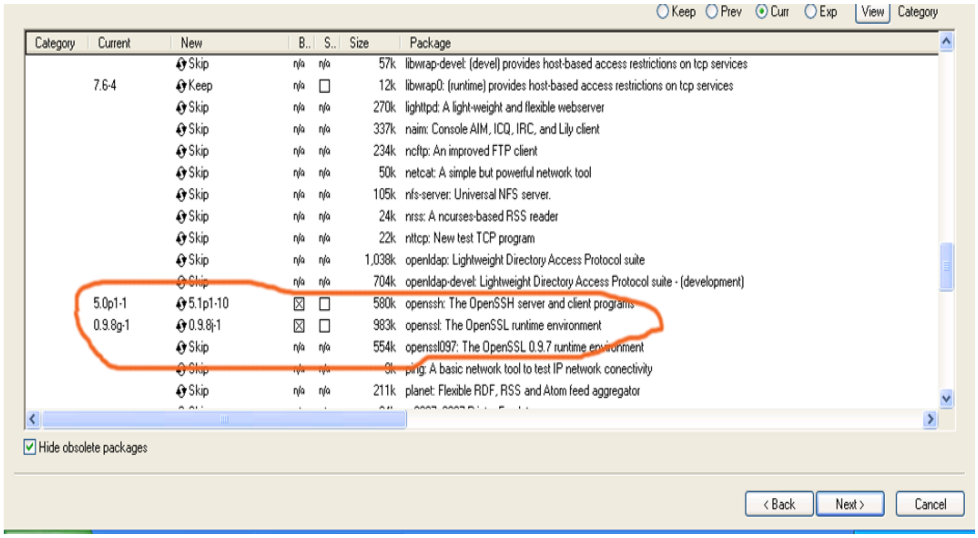

FIGURE A.2 – *Installation de openssh*

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.

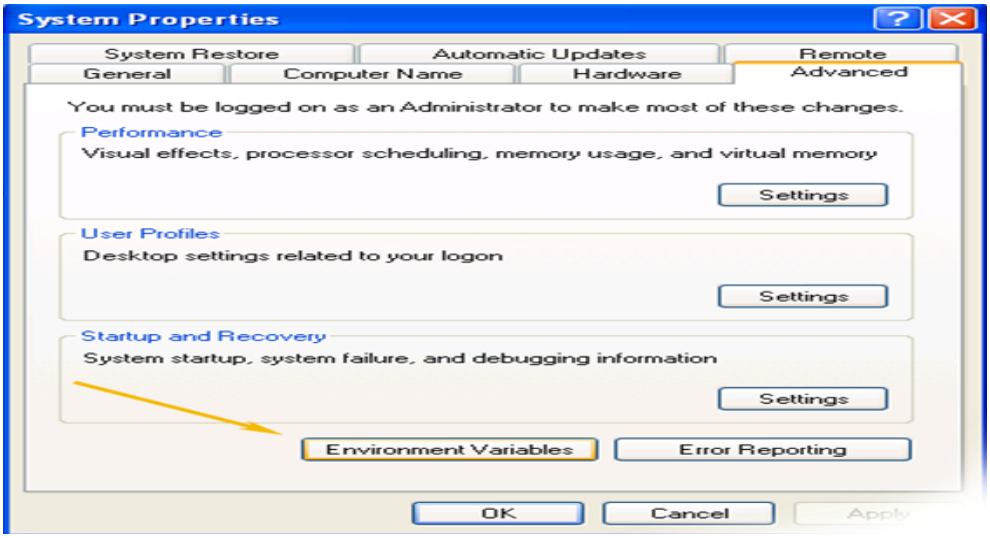

FIGURE A.3 – *Modification de la variable path*

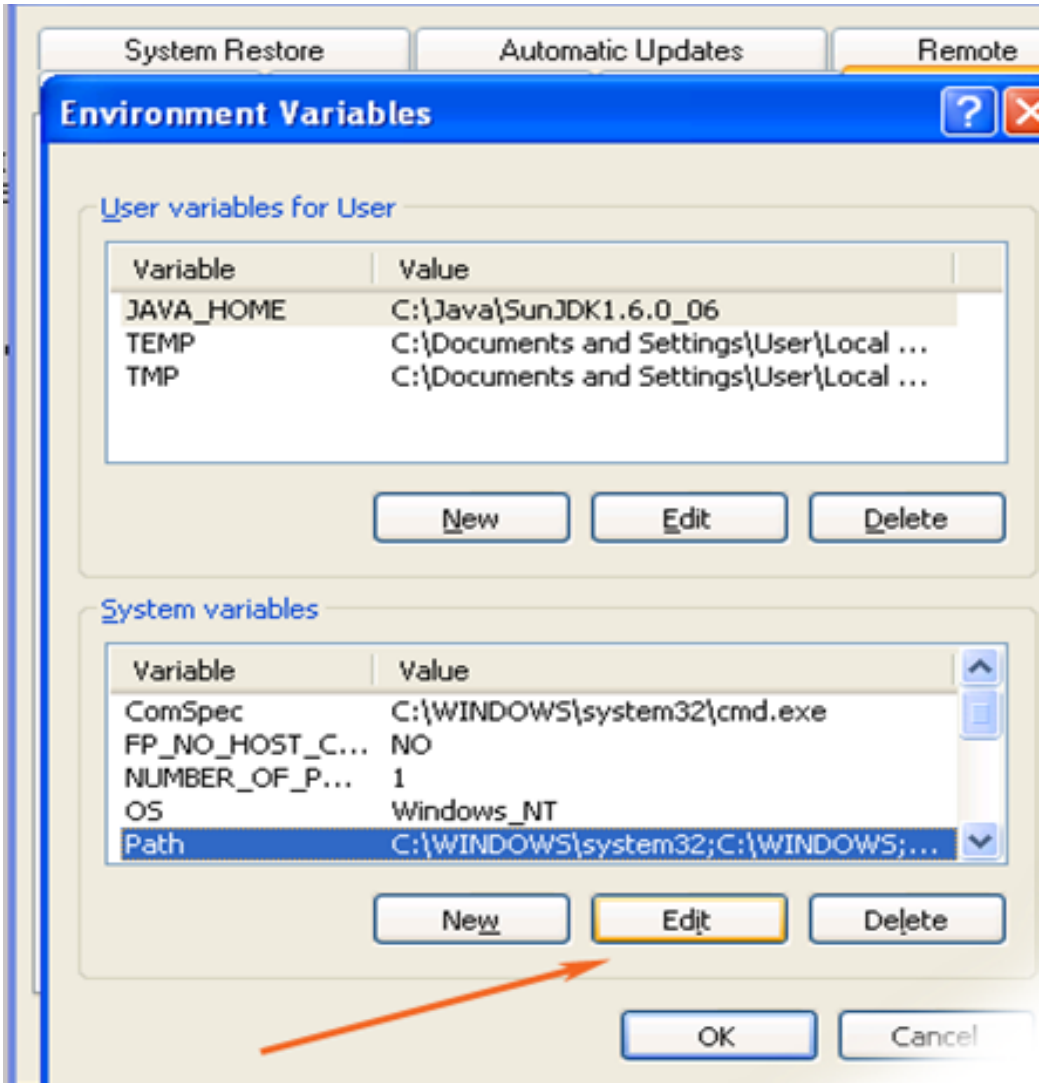

FIGURE A.4 – *Modification de la variable path (suite)*

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.

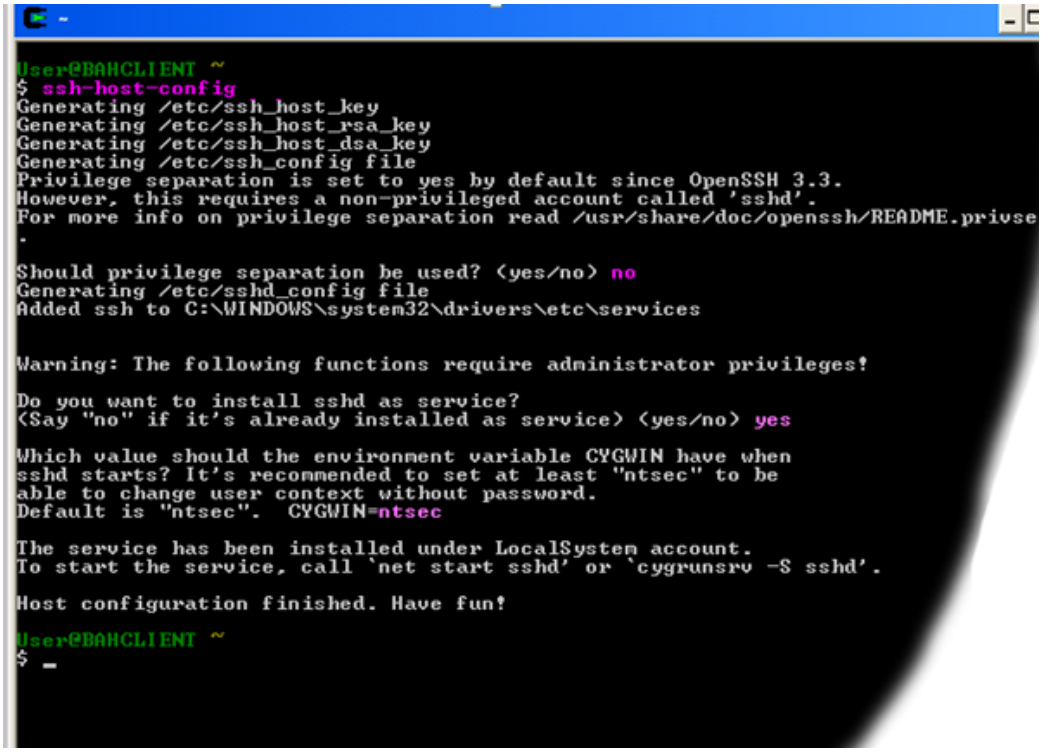

FIGURE A.5 – *Configuration du service ssh*

Démarer 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 disparait et l'état de CYGWIN sshd change.

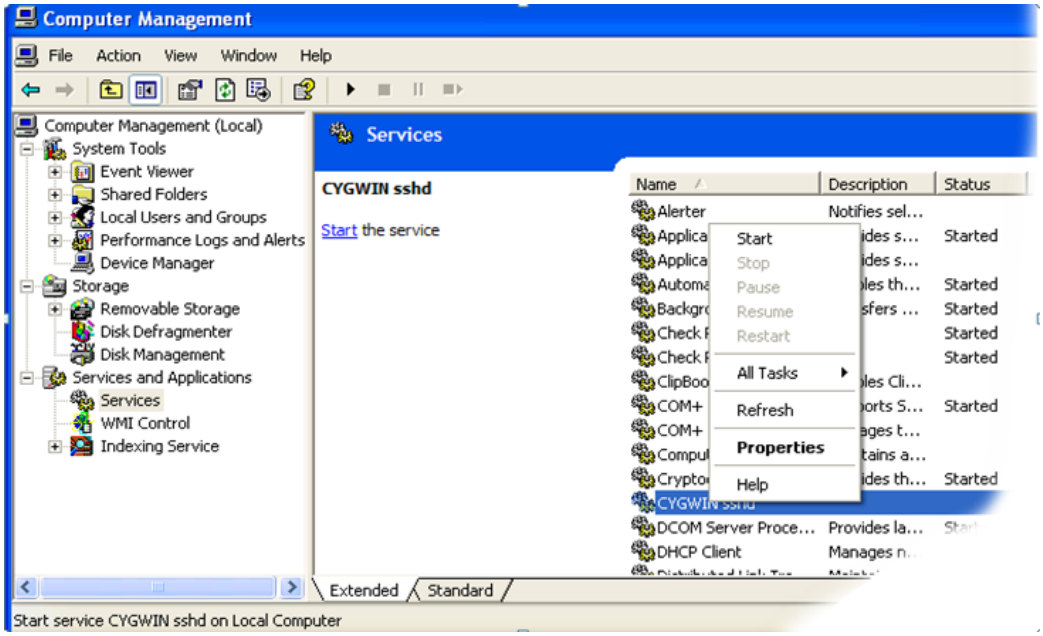

FIGURE A.6 – *Démarer le service ssh*

Configuration de la clé d'authorization

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.

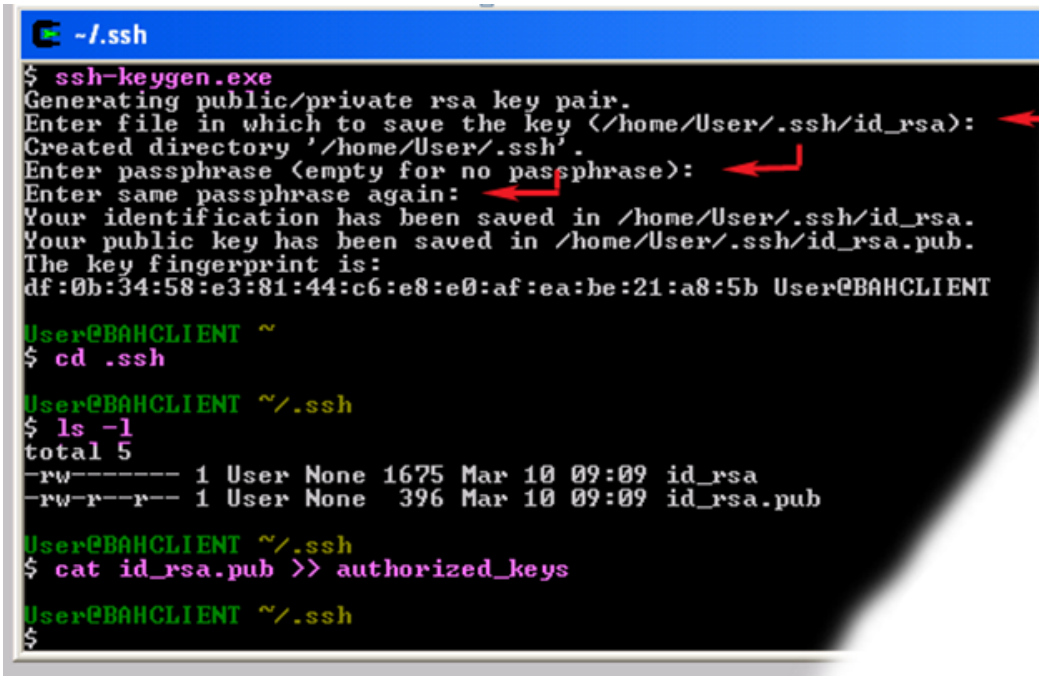

FIGURE A.7 – *Configuration de la clé d'authorization*

— Pour enregistrer les nouvelles clés d'autorisation, entrez la commande suivante (notez que les doubles crochets fortement inclinés ils sont trés 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.

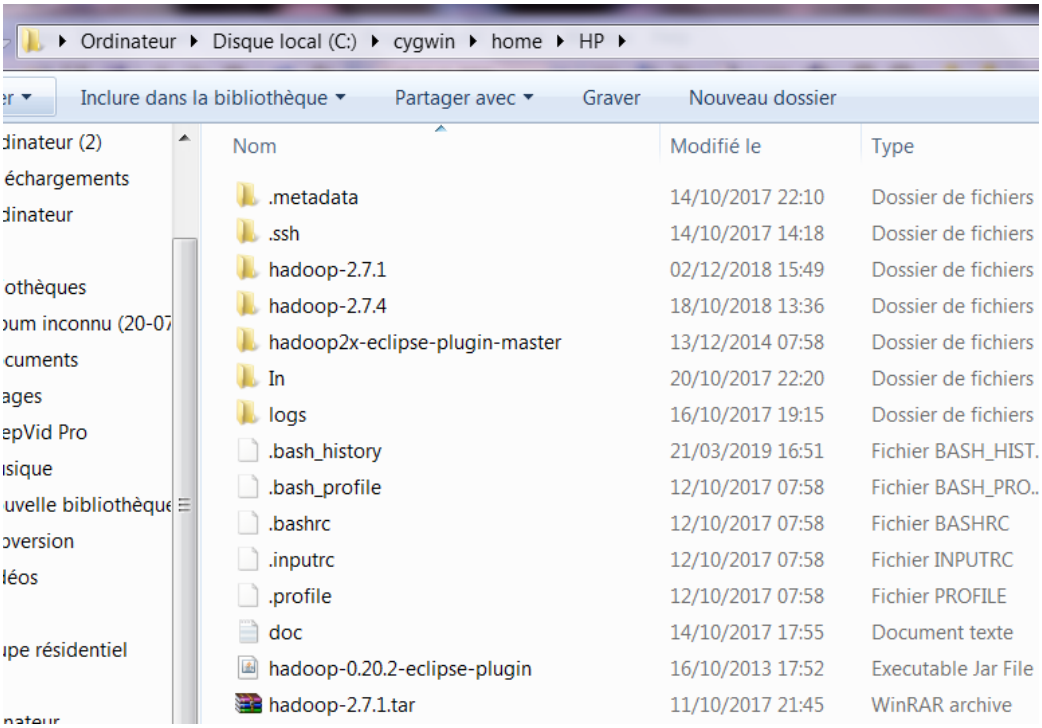

Figure A.8 – *Copier le fichier archive Hadoop*

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.

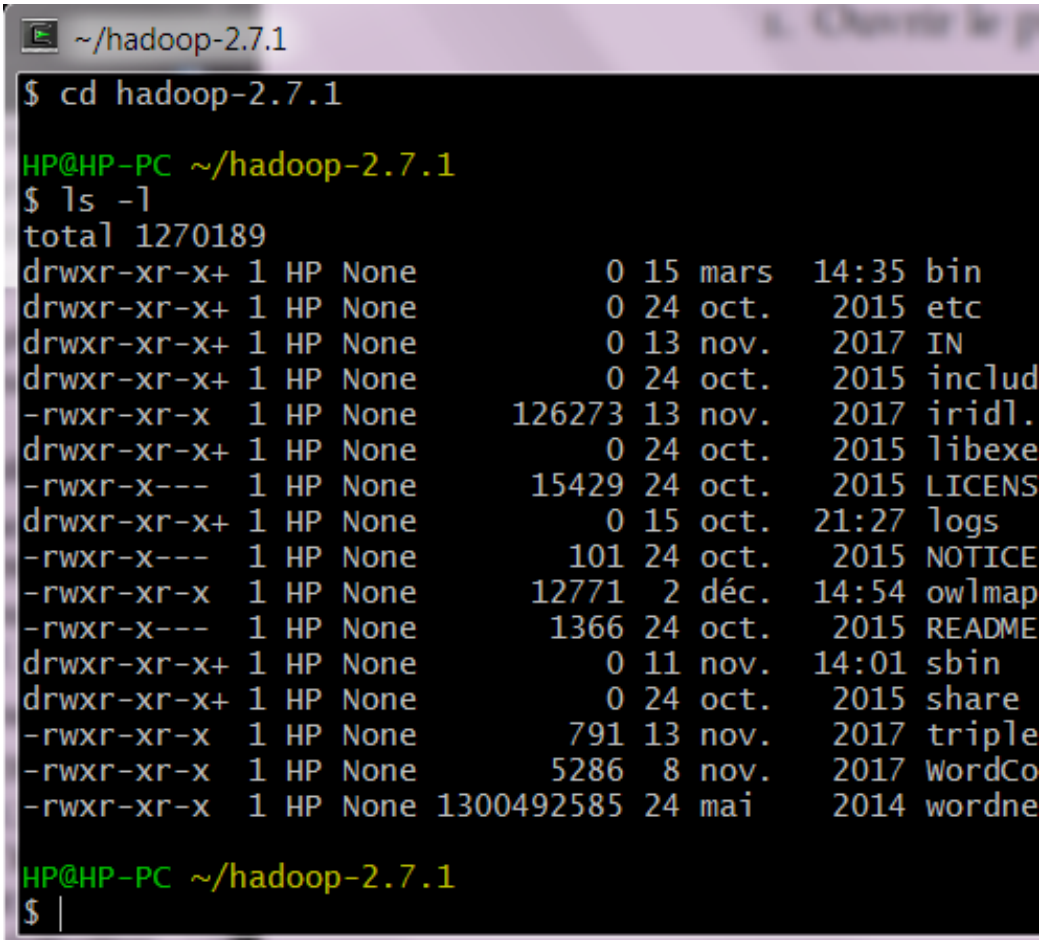

FIGURE A.9 – *Copier le fichier archive Hadoop*

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

```
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>
25
```

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

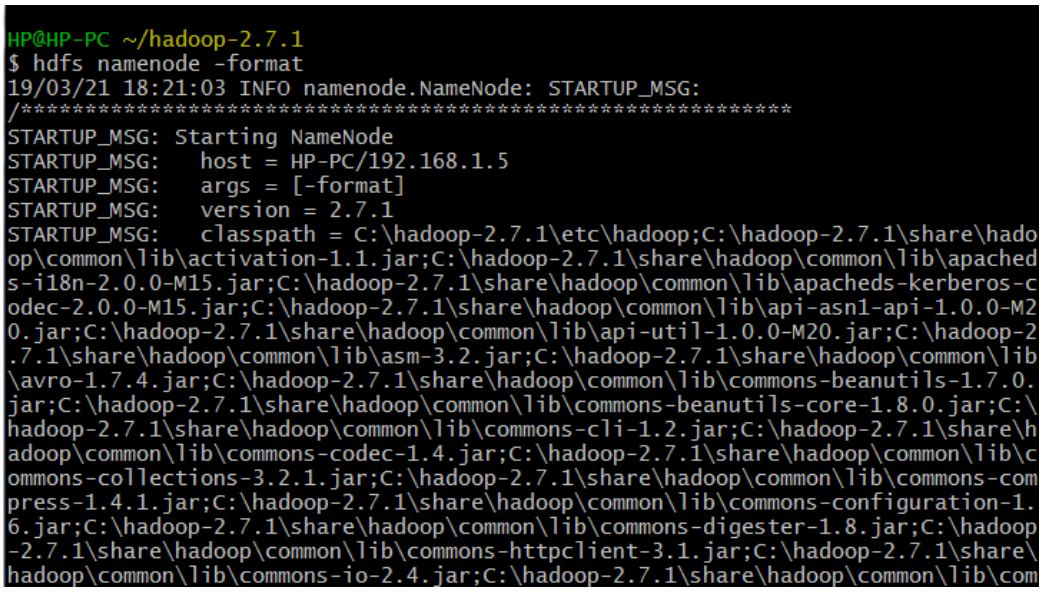

FIGURE A.13 – *Formater HDFS*

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 ficher "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 crée, figure A.19

— d'autres commandes seront testées pendant le TP.

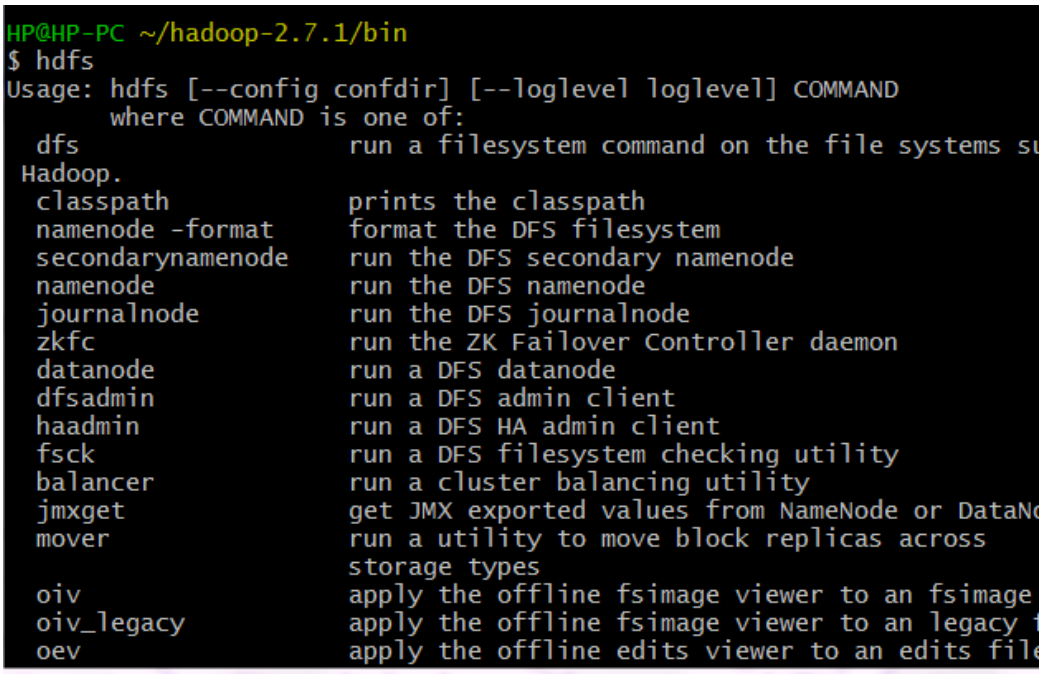

FIGURE A.14 – *Commande hdfs*

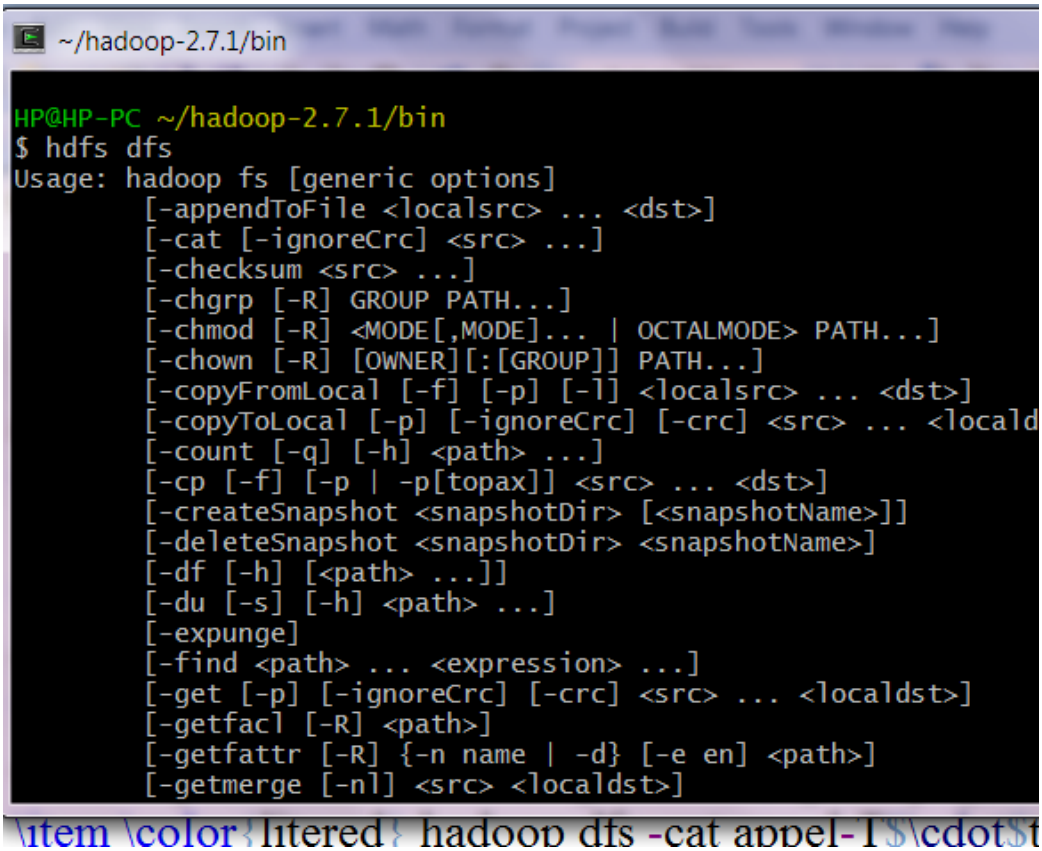

FIGURE A.15 – *Commande hdfs dfs*

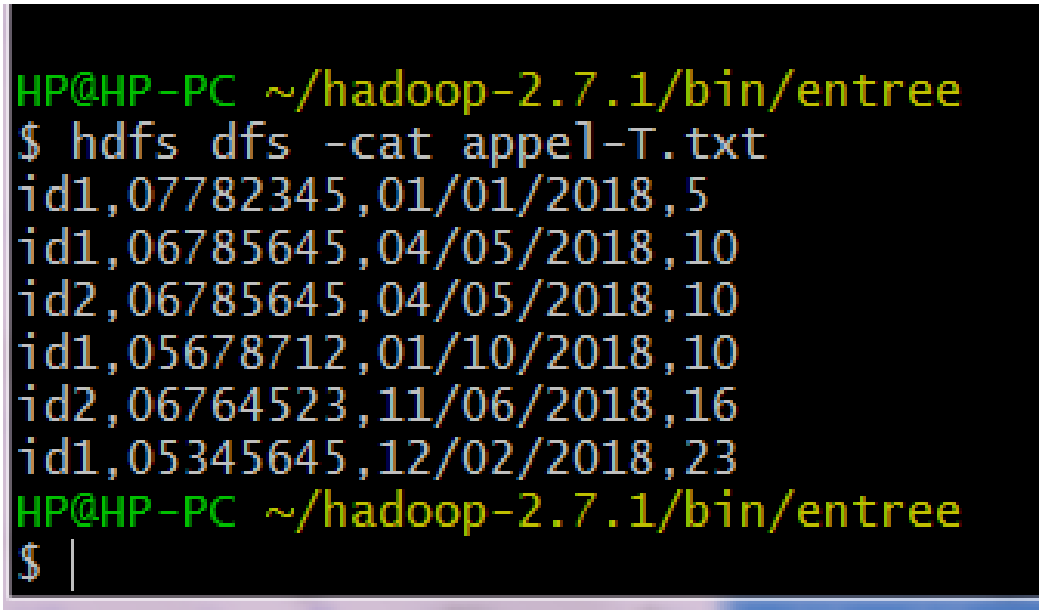

FIGURE A.16 – *Commande hdfs dfs -cat appel-T·txt*

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

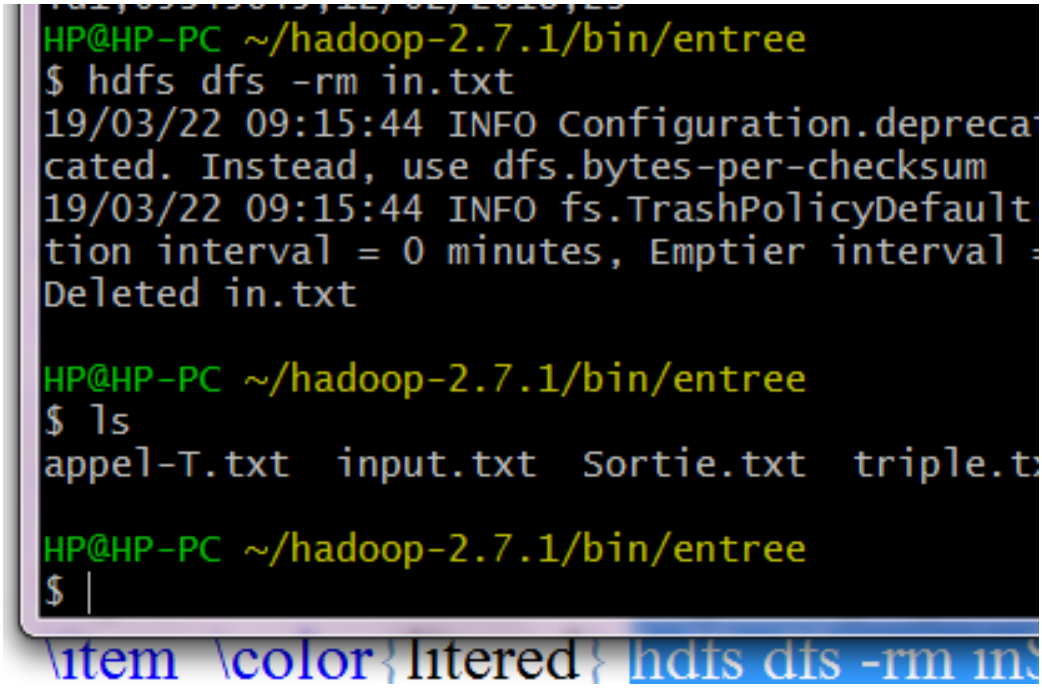

FIGURE A.18 – *Commande hdfs dfs -rm in·txt*

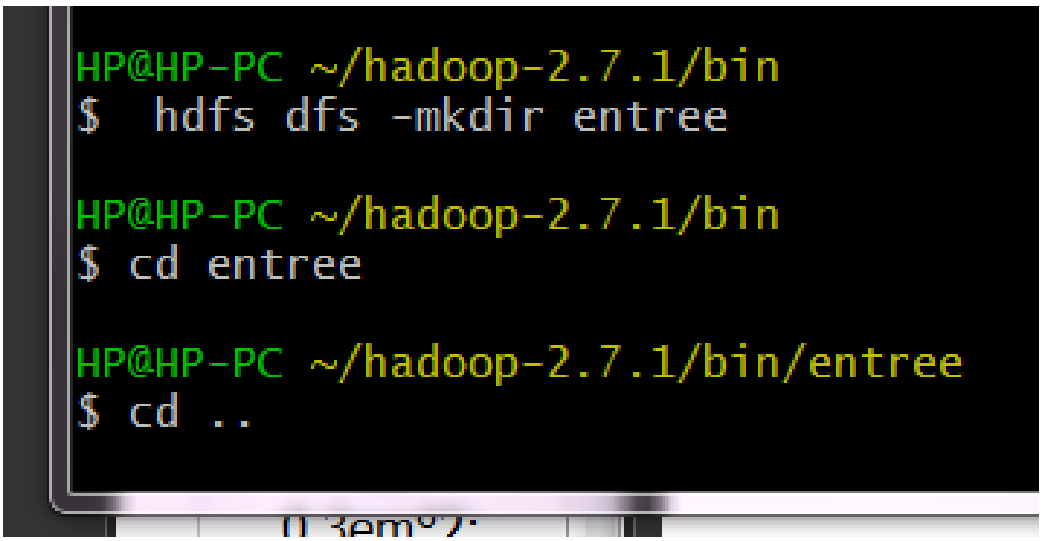

FIGURE A.19 – *Commande hdfs dfs -mkdir entree*

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. Ecrie le progamme de l'exercie1 du chapitre ?? dans la classe, figure A.21 ;
6. Selectionner le File , cliquez sur Export , puis Jar file , selectionnez 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" ;

— Tapez ls ls pour afficher le contenu du répertoire de sortie ;

— Le résultat est sauvegardé dans le fichier "part-r-00000" ;

— Afficher le contenu de ce fichier et verifier le résultat, figure A.25.

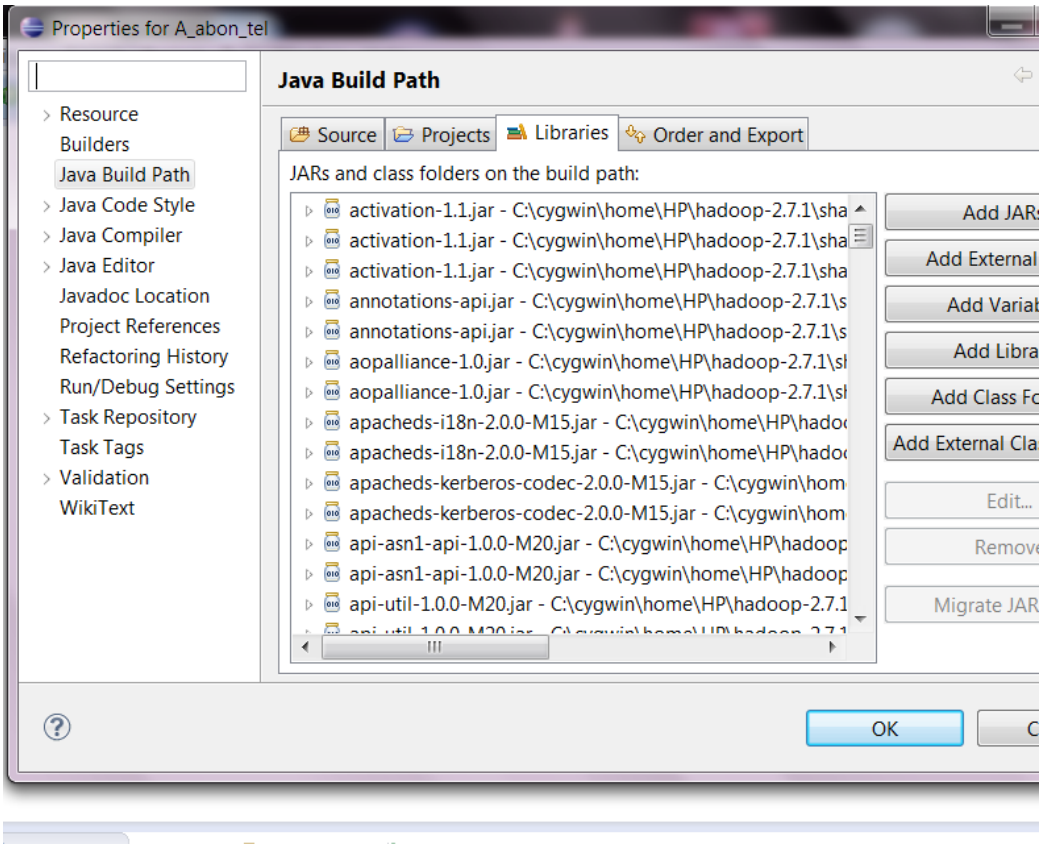

FIGURE A.20 – *Insertion des fichiers Jar Hadoop dans le projet*

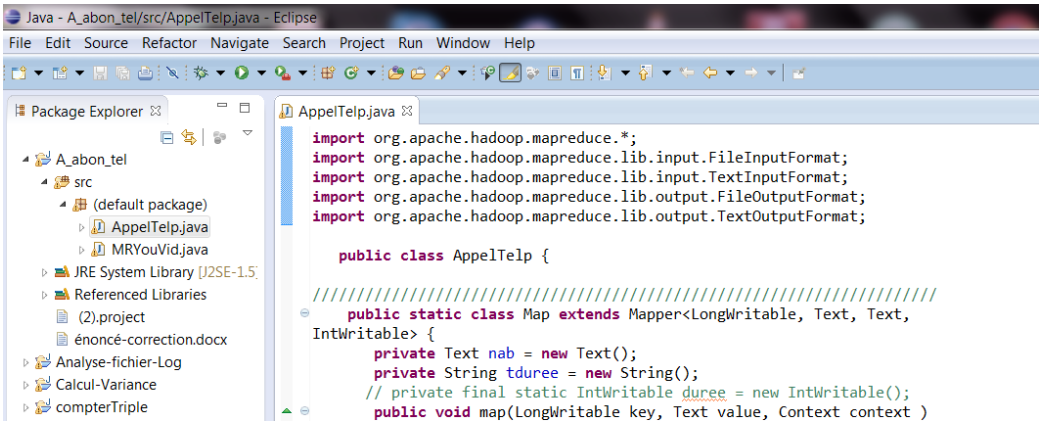

FIGURE A.21 – *Insertion du programme de l'exercie1 du chapitre ?? dans Eclipse*

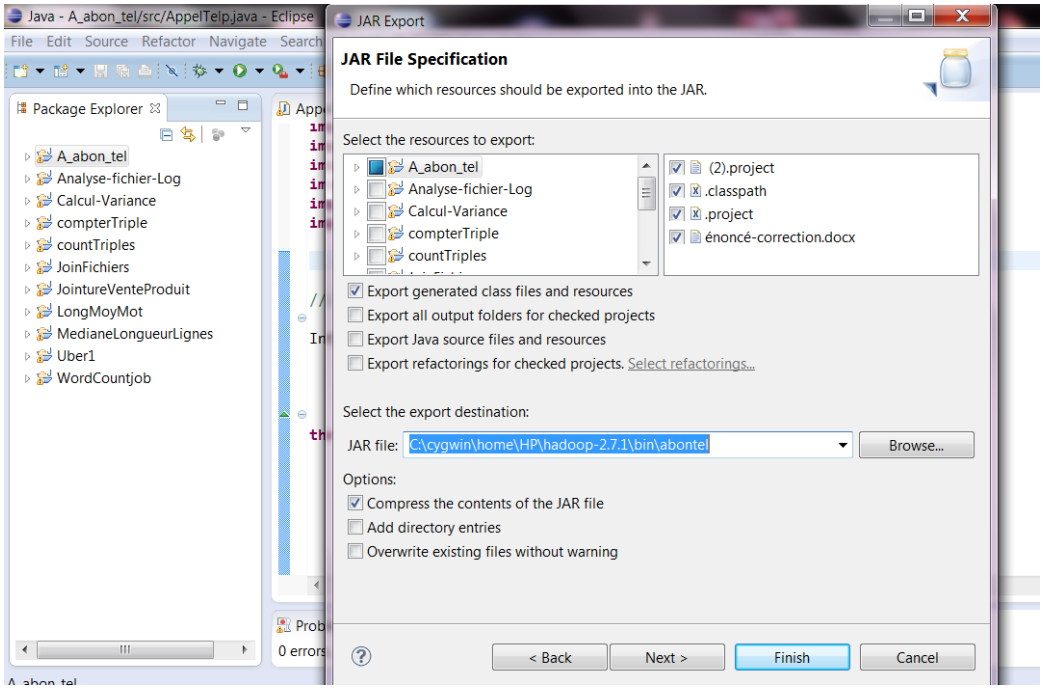

FIGURE A.22 – *Génération du fichier JAR*

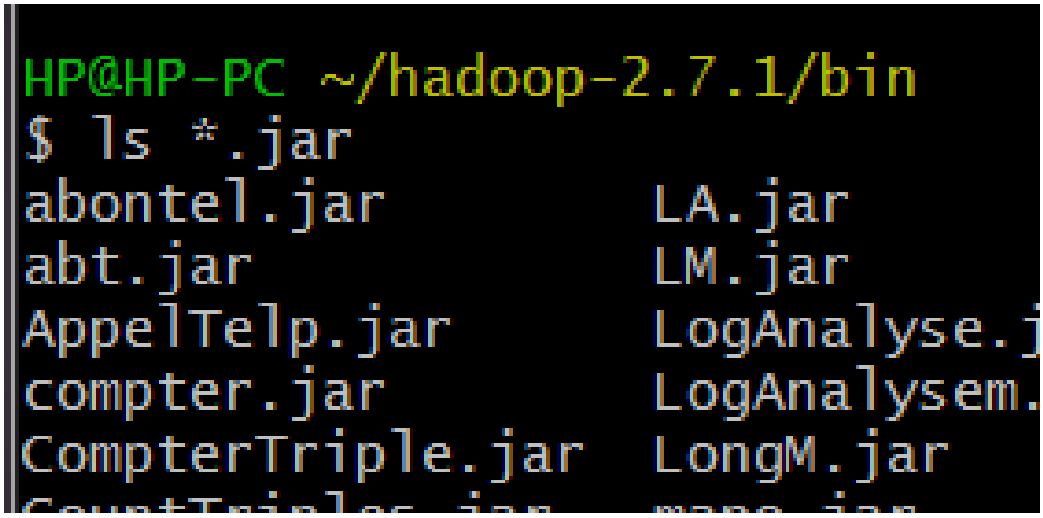

FIGURE A.23 – *Vérification si le programme .jar généré existe*

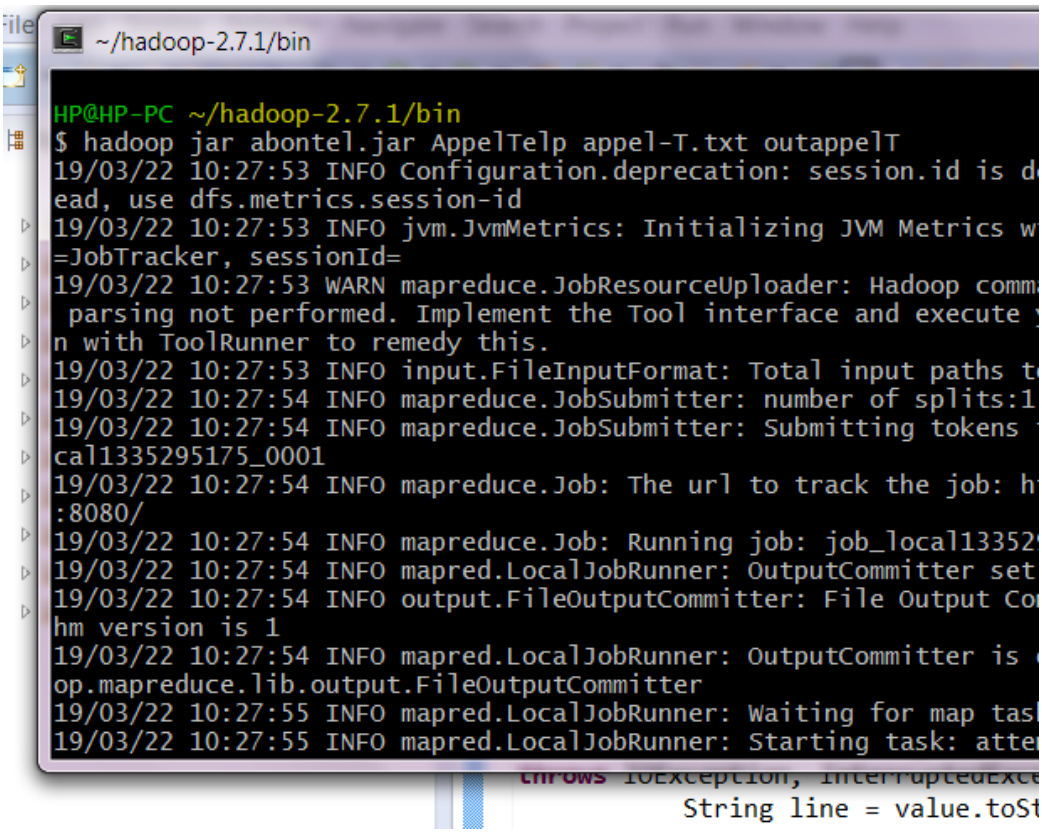

FIGURE A.24 – *Exécution du programme*

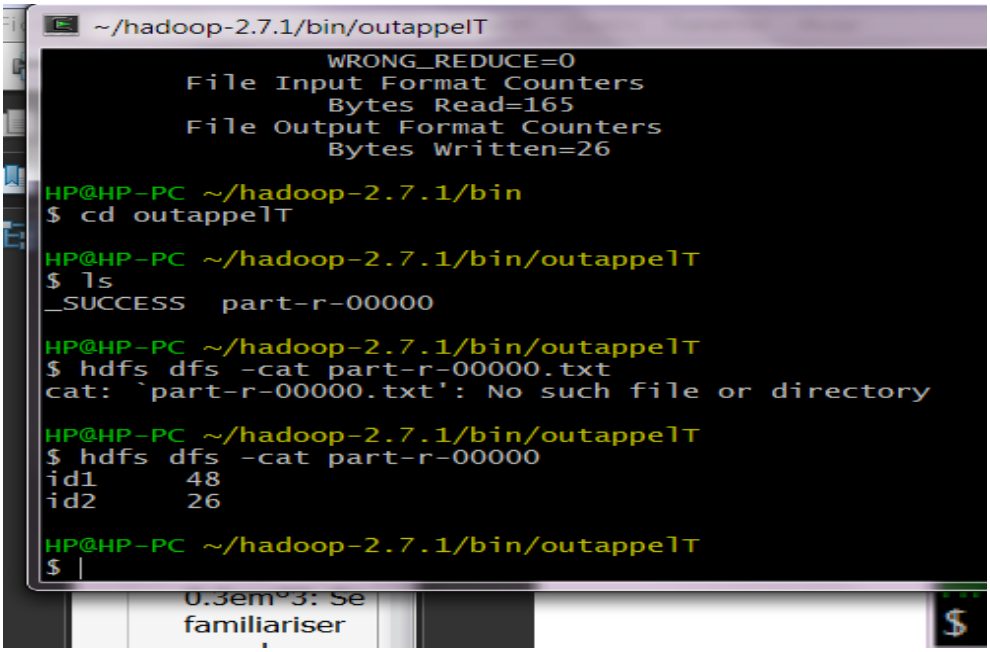

FIGURE A.25 – *Affichage des résultats*

Bibliographie

[01]. site : Le big data, 2019. URL https://www.lebigdata.fr/definition-big-data.

[02]. Stephane Vialle, CentraleSuplec. Big Data : Informatique pour les donnees 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 modele mapreduce dans l'environnement distribue, 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 nÂ°1 : Introduction Ã 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. OReilly, 2012.

[20]. Tom White. Hadoop : The Definitive Guide, 4th Edition. OReilly, 2015.

[21]. Hadoop Training. 2018. URL https://www.slideshare.net/AnandMHadoop/session-19-mapreduce.