# Travaux pratiques: implémentation de Hadoop MapReduce “from scratch” en Java.

## Etape 1: faire un programme séquentiel non parallélisé qui compte le nombre d'occurrences des mots dans un fichier.

**Prérequis pour les questions suivantes:**

* **programmer en Java**[**https://perso.telecom-paristech.fr/bellot/CoursJava/**](https://perso.telecom-paristech.fr/bellot/CoursJava/)
* **manipuler l’environnement de développement Eclipse**[**https://perso.telecom-paristech.fr/bellot/CoursJava/tps/**](https://perso.telecom-paristech.fr/bellot/CoursJava/tps/)
* **lire et écrire dans un fichier en java**[**https://perso.telecom-paristech.fr/bellot/CoursJava/JavaIO.html**](https://perso.telecom-paristech.fr/bellot/CoursJava/JavaIO.html)

#### Premier comptage en séquentiel pur

Implémentez un logiciel en java qui compte le nombre d’occurrences des mots d’un fichier d’entrée de manière non parallélisée (monothread, une seul thread), en utilisant un seul processeur.

Quelle structure de donnée est la plus pertinente pour stocker les résultats: List, HashMap ou HashSet ou une autre ? Pour quelle raison ?

Testez votre programme avec un fichier d’entrée *input.txt* avec comme contenu:  
  
Deer Beer River  
Car Car River  
Deer Car Beer

Résultat:

Deer 2

Beer 2

River 2

Car 3

#### Premier tri en séquentiel pur

Modifiez votre programme pour trier par nombre d'occurrences:

Résultat:

Car 3

Deer 2

Beer 2

River 2

#### Deuxième tri alphabétique en séquentiel pur

Modifiez le programme pour trier alphabétiquement pour les mots à égalité du nombre d’occurrences:

Résultat:

Car 3

Beer 2

Deer 2

River 2

#### Test du programme séquentiel sur le code forestier de Mayotte

Testez ensuite votre programme avec le code forestier de Mayotte disponible sur github *forestier\_mayotte.txt* :   
<https://github.com/legifrance/Les-codes-en-vigueur>

Votre programme a-t-il fonctionné du premier coup ?

Vérifiez en ouvrant le fichier texte qu’il contient bien du texte et non du code HTML.

Ne perdez pas de temps à corriger les éventuelles erreurs dues aux caractères spéciaux ou à des mots suspects ou illisibles (de toutes façons par la suite il y aura du chinois dans le texte).

#### Les 50 mots du code de la déontologie de la police nationale

Testez votre programme avec le code de déontologie de la police nationale disponible sur github *deontologie\_police\_nationale.txt* : <https://github.com/legifrance/Les-codes-en-vigueur>

De même ne perdez pas de temps à filtrer les caractères spéciaux ou autres mots bizarres. Pourquoi ? Car nous travaillerons ensuite sur des textes en chinois, japonais, arabe et d’autres langues. Si vous implémentez une étape de filtrage ici en français elle ne servira à rien par la suite. Quels sont les 5 premiers mots (qui ressemblent à des mots) parmi les 50 premiers de la liste triée résultat ? Gardez la réponse pour l’intégrer au rapport.

police : 38

article : 20

nationale : 20

titre : 13

fonctionnaires : 10

#### Les 50 mots du code du domaine public fluvial

Testez votre programme avec le code du domaine public fluvial *domaine\_public\_fluvial.txt*.

Quels sont les 5 premiers mots (qui ressemblent à des mots) parmi les 50 premiers de la liste triée résultat ? Gardez la réponse pour l’intégrer au rapport.

article : 107

bateau : 106

tribunal : 72

lieu : 69

larticle : 68

dimmatriculation : 64

#### Les 50 mots du code de la santé publique

#### Testez votre programme avec le code de la santé publique *sante\_publique.txt*.

Quels sont les 5 premiers mots (qui ressemblent à des mots) parmi les 50 premiers de la liste triée résultat ? Gardez la réponse pour l’intégrer au rapport.

article : 23846

larticle : 22998

santé : 20197

directeur : 8604

lagence : 8290

conditions : 8096

#### Chronométrage du programme séquentiel

Chronométrer votre programme sur le code de la santé publique.

Chronométrage possible avec:  
long startTime = System.currentTimeMillis();  
...  
long endTime   = System.currentTimeMillis();  
long totalTime = endTime - startTime;

Combien de temps faut-il pour chacune des étapes:

* Compter le nombre d’occurrences
* Tri (par nombre d'occurrences et alphabétique)

Gardez la réponse pour l’intégrer au rapport.

Counting time is:2886

Sorting time is:34

#### Travailler sur des plus gros fichiers

Testez votre programme sur un cas réel: un extrait de toutes les pages internet transformées au format texte brut (format WET). Toutes les pages sur  internet au format texte sont disponibles sur <http://commoncrawl.org/the-data/get-started/> : chaque mois, environ 3 milliards de pages web, soit 250 To de données sont stockées. Ces données sont disponibles par tranche de moins d’1Go environ, vous travaillerez sur une tranche de 380Mo.

J’ai choisi une tranche en particulier pour avoir une comparaison entre nous (vous pouvez tester sur d’autres tranches si vous voulez). Téléchargez cette tranche ici:

<https://commoncrawl.s3.amazonaws.com/crawl-data/CC-MAIN-2017-13/segments/1490218189495.77/wet/CC-MAIN-20170322212949-00140-ip-10-233-31-227.ec2.internal.warc.wet.gz>

Décompressez et obtenez le fichier CC-MAIN-20170322212949-00140-ip-10-233-31-227.ec2.internal.warc.wet

Il s’agit d’une tranche contenant un ensemble de sites internet au format texte brut (WET).

Testez votre programme avec ce fichier en entrée. Chronométrez-le.

Gardez la réponse pour l’intégrer au rapport.

#### **Préparez un document de travail qu’il faudra rendre à la fin de l’unité d’enseignement.**

Créez un document dans lequel vous allez expliquer votre implémentation au fur et à mesure. Vous obtiendrez également par la suite (dans les prochaines étapes) les mesures de chronométrage et calculerez le speedup (l’accélération) et la portion de code parallèle (le taux de parallélisation) à chaque fois avec vos nouvelles mesures. Pour comparer deux mesures, il faudra bien faire attention d’avoir le même jeu de données en entrée et le même résultat en sortie... Vous devez prouver que vous obtenez le même résultat en sortie en utilisant un autre logiciel (qui fait de la recherche du nombre de mots) et piocherez au hasard des mots pour savoir si dans le cas séquentiel et dans le cas réparti vous avez les même résultats.

Etape 2: travailler avec plusieurs ordinateurs en réseau.

**Prérequis et documentation pour les questions suivantes :**

* **avoir un compte et pouvoir se connecter aux machines de l’école (contacter la DSI de l’école dans le cas contraire)**
* **lancer un interpréteur de commande (console linux , shell) pour taper des commandes:**[**https://fr.wikipedia.org/wiki/Interpréteur\_de\_commandes**](https://fr.wikipedia.org/wiki/Interpréteur_de_commandes)  
  [**https://fr.wikipedia.org/wiki/Shell\_Unix**](https://fr.wikipedia.org/wiki/Shell_Unix)
* **connaître quelques commandes de base sous Linux**[**https://fr.wikipedia.org/wiki/Commandes\_Unix**](https://fr.wikipedia.org/wiki/Commandes_Unix)
* [**https://prod-edx-mktg-edit.edx.org/course/linux-basics-the-command-line-interface**](https://prod-edx-mktg-edit.edx.org/course/linux-basics-the-command-line-interface)
* **savoir qu’un ordinateur a un nom d’hôte (hostname) et plusieurs adresses IP:**[**https://en.wikipedia.org/wiki/Hostname**](https://en.wikipedia.org/wiki/Hostname)  
  [**https://fr.wikipedia.org/wiki/Adresse\_IP**](https://fr.wikipedia.org/wiki/Adresse_IP)
* **savoir qu’un ordinateur peut faire partie d’un domaine (comme à l’école, le domaine enst.fr ou le domaine telecom-paristech.fr )**[**https://fr.wikipedia.org/wiki/Nom\_de\_domaine**](https://fr.wikipedia.org/wiki/Nom_de_domaine)
* **savoir qu’un nom peut être transformé en adresse IP (et inversement) par un serveur qui gère le système du nom de domaine (DNS, Domain Name System)**[**https://fr.wikipedia.org/wiki/Domain\_Name\_System**](https://fr.wikipedia.org/wiki/Domain_Name_System)
* **se connecter à distance à un ordinateur avec SSH en ligne de commande**[**https://fr.wikipedia.org/wiki/Secure\_Shell**](https://fr.wikipedia.org/wiki/Secure_Shell)  
  [**http://www.commentcamarche.net/faq/74-se-connecter-a-distance-avec-ssh-linux**](http://www.commentcamarche.net/faq/74-se-connecter-a-distance-avec-ssh-linux)

1. **Nom court, nom long**

Quel est le nom COURT de votre ordinateur (le nom simple sans le domaine) ?   quel est le nom LONG de votre ordinateur (le nom avec le domaine) ? Comment les connaître en ligne de commande ? Sur les ordinateurs de l’école, est-il possible d’obtenir ces noms autrement qu’en ligne de commande ? Ajoutez les réponses à votre rapport.

#### Adresse ip

Comment connaître les adresses (plusieurs) IP de votre ordinateur en ligne de commande ? Autrement (en passant par un site internet par exemple) ? Ajoutez les réponses à votre rapport.

#### Du nom vers l’IP

Comment à partir du nom d’un ordinateur, obtenir les adresses IP en ligne de commande ? Ajoutez les réponses à votre rapport.

#### De l’IP vers le nom

Comment, à partir d’une adresse IP, obtenir les noms associés en ligne de commande ? Ajoutez les réponses à votre rapport.

#### Ping pong à l’intérieur!

Testez la communication avec d’autres ordinateurs (pas le vôtre) depuis le réseau de l’école en utilisant la commande ping (pour arrêter le ping faire CTRL + C). suivi du nom court, du nom long, de l’IP. Les trois méthodes fonctionnent-elles ? Ajoutez les réponses à votre rapport.

Ping adresse

#### Ping pong à l’extérieur

Si vous effectuez le ping depuis un réseau différent, il est possible que celui ne fonctionne pas (filtrage des accès vers le réseau de l’école depuis un réseau extérieur), contactez la DSI pour mettre en place une connection VPN / OpenVPN afin d’être sur le même réseau que les machines en salle de TP.

#### Calculer en ligne de commande sur l’ordinateur local

Comment lancer un calcul en ligne de commande sur votre ordinateur (par exemple 2 + 3) ? Parmi les multiples réponses possibles, lesquelles permettent de lancer le calcul et d’obtenir le résultat en appuyant une seule fois sur la touche <Entrée> ? Ajoutez les réponses à votre rapport.

Calc 3+5

#### Calculer en ligne de commande sur un ordinateur distant

Comment lancer un calcul  (par exemple 2 + 3) en ligne de commande sur un autre ordinateur (à distance) ? Il faudra certainement vous authentifier avec un mot de passe. Comment obtenir le résultat du calcul immédiatement après avoir tapé son mot de passe ? Ajoutez les réponses à votre rapport.

Ssh adresse calc 3+5

#### Calculer à distance sans mot de passe

Comment lancer un calcul à distance en utilisant SSH sans taper le mot de passe et en une seule ligne de commande (c’est à dire qu’on appuie sur <Entrée> et on a le résultat directement)?

Attention: en utilisant vos ordinateurs personnels( a priori ce n’est pas le cas sur les ordinateurs de l’école), il faut parfois vérifier le “fingerprint” avec le message:

The authenticity of host 'mint.phcomp.co.uk (78.32.209.33)' can't be established.  
RSA key fingerprint is 6a:de:e0:af:56:f8:0c:04:11:5b:ef:4d:49:ad:09:23.  
Are you sure you want to continue connecting (yes/no)? yes

Pour ne plus avoir à rentrer quoi que ce soit dans le terminal, il faut donc désactiver la vérification de ce “fingerprint”, sinon vous aurez l’erreur

Host key verification failed

Suivez ce guide pour désactiver la vérification de “fingerprint”:

<https://www.shellhacks.com/disable-ssh-host-key-checking/>

Information poussé sur le fichier :

df -k nomfichier

Depuis A copier un fichier de B vers C :

scp c125-08:adresseDuFichier c125-11 :adresseDeEcritureFichier

scp -r dossier lieu

ss