YARN & MapReduce 2

Lucas BAKALIAN & Michaël HATOUM

August 15, 2021

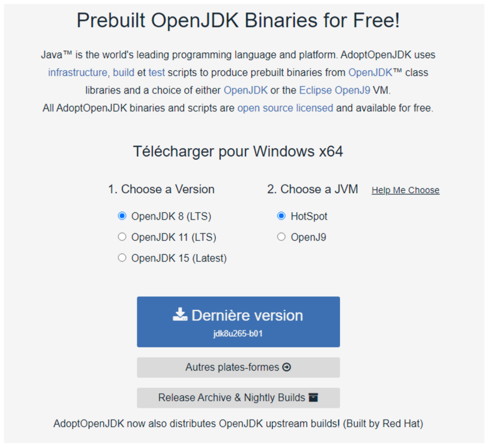
1 MapReduce JAVA

In this tutorial, we will use write MapReduce jobs using JAVA. Before that, we will install prerequirements, if you already have them, skip thoses parts.

1.1 Install OpenJDK 8

1.1.1 Microsoft Windows

Download and install OpenJDK for Windows from here.



Prebuilt OpenJDK Binaries for Free!

Java™ is the world's leading programming language and platform. AdoptOpenJDK uses infrastructure, build et test scripts to produce prebuilt binaries from OpenJDK™ class libraries and a choice of either OpenJDK or the Eclipse OpenJ9 VM.
All AdoptOpenJDK binaries and scripts are open source licensed and available for free.

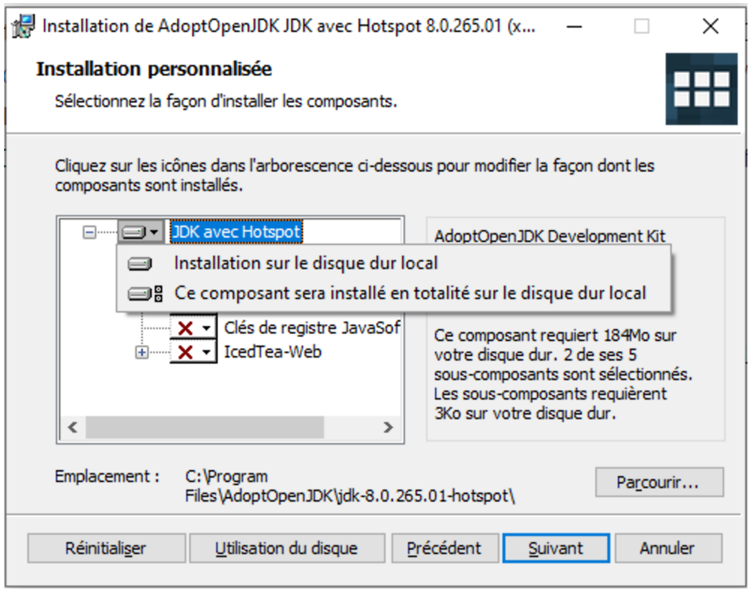
Télécharger pour Windows x64

1. Choose a Version: [✓ OpenJDK 8 (LTS)] OpenJDK 11 (LTS) OpenJDK 15 (Latest)

2. Choose a JVM: [✓ HotSpot] OpenJ9

AdoptOpenJDK now also distributes OpenJDK upstream builds! (Built by Red Hat)

Start the installer and configure it to install everything:



Installation de AdoptOpenJDK JDK avec Hotspot 8.0.265.01 (x...

Installation personnalisée

Sélectionnez la façon d'installer les composants.

Cliquez sur les icônes dans l'arborescence ci-dessous pour modifier la façon dont les composants sont installés.

JDK avec Hotspot
- Installation sur le disque dur local
- Ce composant sera installé en totalité sur le disque dur local
- Clés de registre JavaSof
- IcedTea-Web

AdoptOpenJDK Development Kit

Ce composant requiert 184Mo sur votre disque dur. 2 de ses 5 sous-composants sont sélectionnés. Les sous-composants requièrent 3Ko sur votre disque dur.

Emplacement : C:\Program Files\AdoptOpenJDK\jdk-8.0.265.01-hotspot\

1.1.2 Mac OS X

It is easier to install OpenJDK on OS X:

```
$ brew tap AdoptOpenJDK/openjdk
$ brew cask install adoptopenjdk8
```

1.2 Git

1.2.1 Microsoft Windows

Download and install 64-bit Git for Windows Setup from here.

1.2.2 Mac OS X

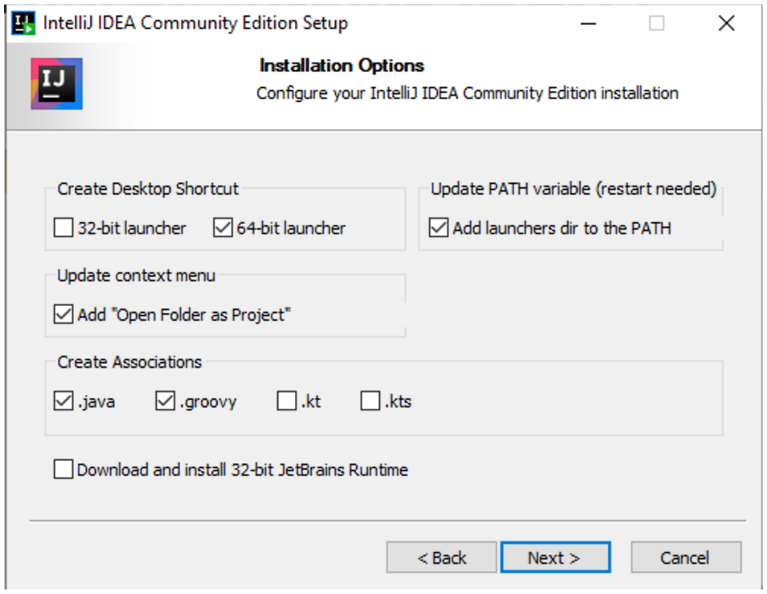
There are several ways to install Git on a Mac. The easiest is probably to install the Xcode Command Line Tools. On Mavericks (10.9) or above you can do this simply by trying to run git from the Terminal the very first time.

```
$ git --version
```

1.3 IntelliJ IDEA

1.3.1 Microsoft Windows

Download and install IntelliJ IDEA for Windows (Community edition) from here. Here the installation options and restart your computer when asked.



IntelliJ IDEA Community Edition Setup

Installation Options

Configure your IntelliJ IDEA Community Edition installation

Create Desktop Shortcut
- [ ] 32-bit launcher
- [✓] 64-bit launcher

Update PATH variable (restart needed)
- [✓] Add launchers dir to the PATH

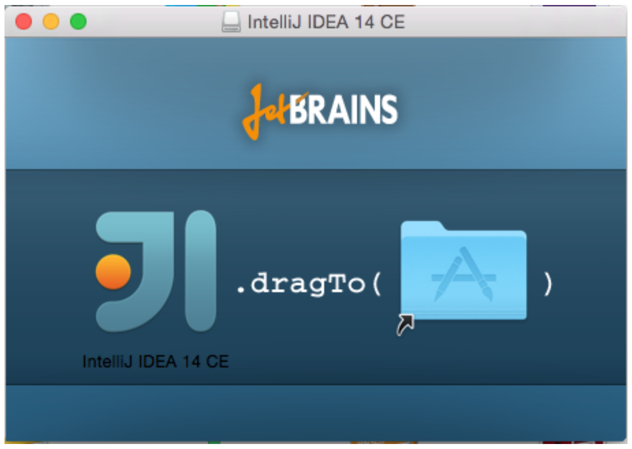
Update context menu
- [✓] Add "Open Folder as Project"

Create Associations
- [✓] .java
- [✓] .groovy
- [ ] .kt
- [ ] .kts

- [ ] Download and install 32-bit JetBrains Runtime

1.3.2 Mac OS X

Download and install IntelliJ IDEA for Mac OSX (Community edition) from here. Open the .dmg folder and move the IDEA app into your applications folder.

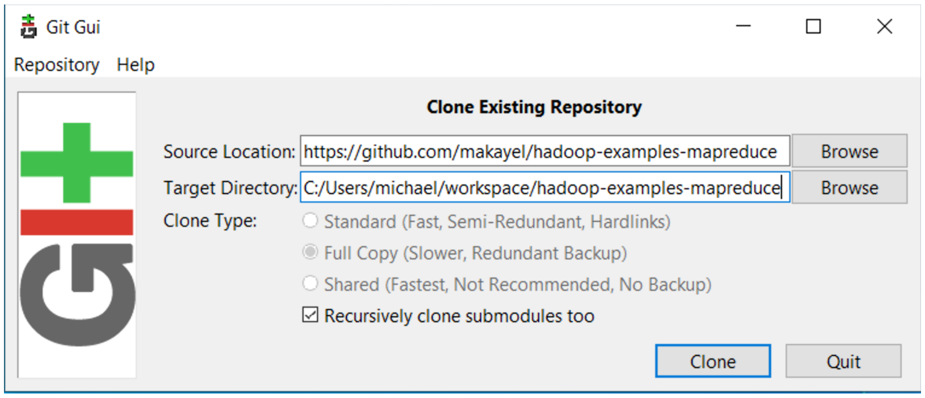


1.4 Clone the project

Here is a JAVA project hosted on github to start your lab.

1.4.1 Microsoft Windows

You can use Git GUI for cloning the repository.



Clone Existing Repository

Source Location: https://github.com/makayel/hadoop-examples-mapreduce

Target Directory: C:/Users/michael/workspace/hadoop-examples-mapreduce

Clone Type:
- ○ Standard (Fast, Semi-Redundant, Hardlinks)
- ● Full Copy (Slower, Redundant Backup)
- ○ Shared (Fastest, Not Recommended, No Backup)

☑ Recursively clone submodules too

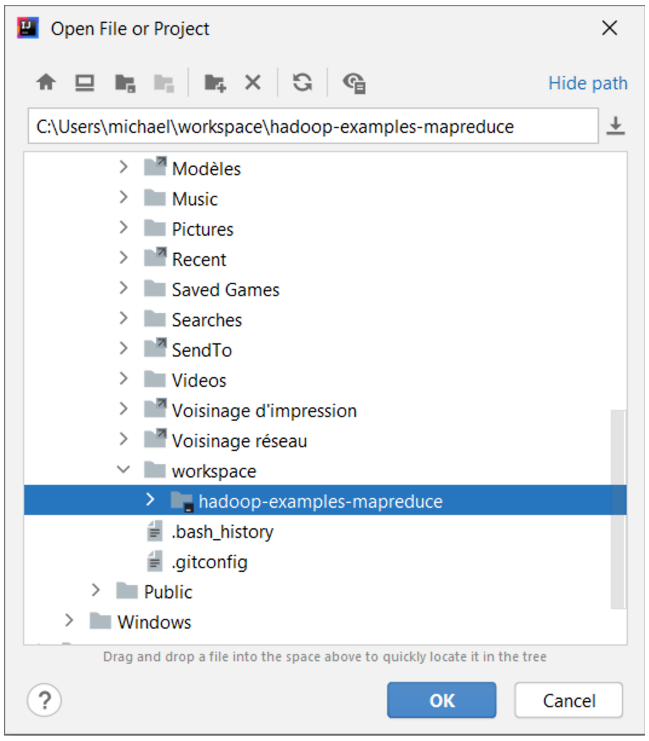
1.4.2 Mac OS X

Use CLI to clone the repository.

```
$ cd /Users/michael/workspace/IdeaProjects
$ git clone https://github.com/makayel/hadoop-examples-mapreduce
```

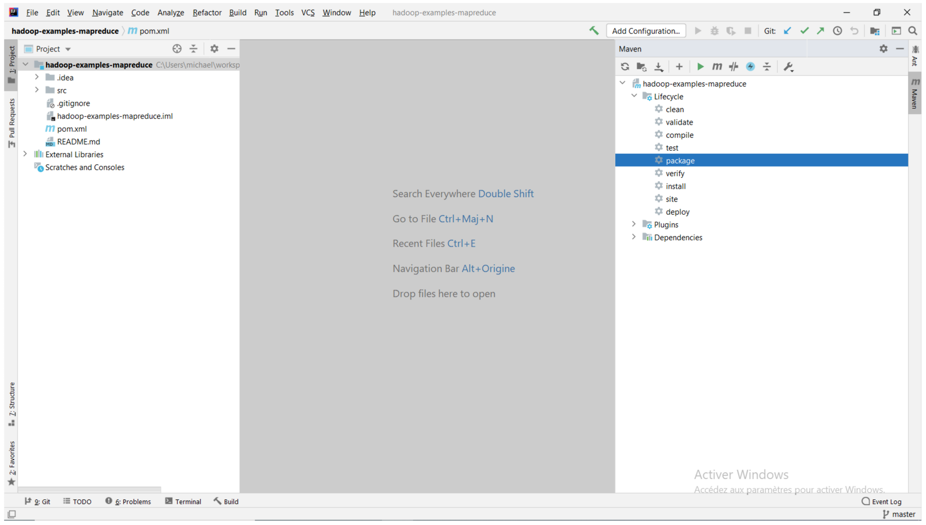
1.5 Import the project

Open a new project (import) from the source cloned.

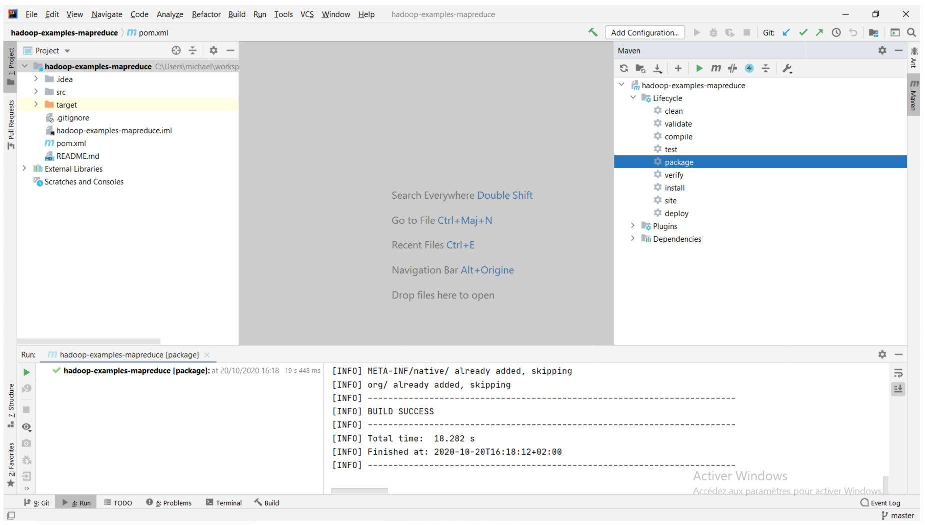


C:\Users\michael\workspace\hadoop-examples-mapreduce

Generate the JAR using maven lifecycle `package`.

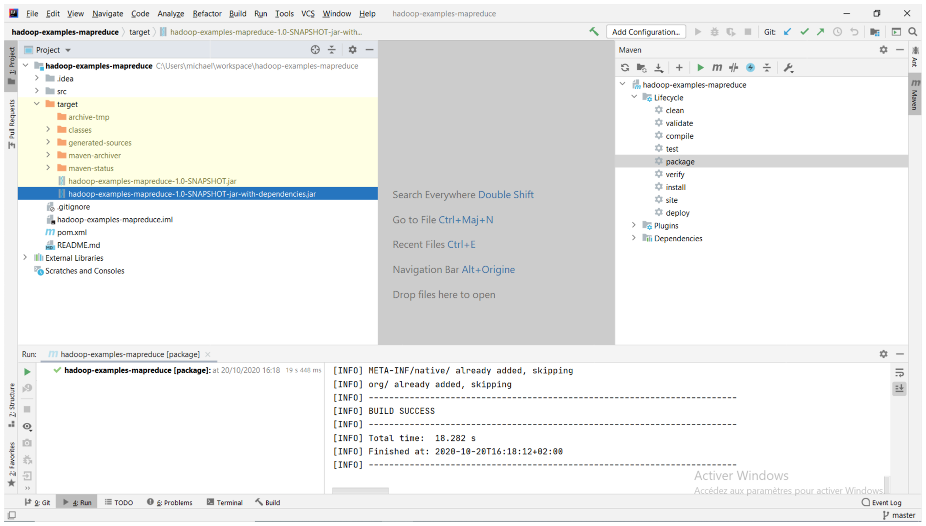

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help hadoop-examples-mapreduce
hadoop-examples-mapreduce pom.xml
Add Configuration...
Project
hadoop-examples-mapreduce C:\Users\michael\worksp
.idea
src
.gitignore
hadoop-examples-mapreduce.iml
pom.xml
README.md
External Libraries
Scratches and Consoles
Search Everywhere Double Shift
Go to File Ctrl+Maj+N
Recent Files Ctrl+E
Navigation Bar Alt+Origine
Drop files here to open
Maven
hadoop-examples-mapreduce
Lifecycle
clean
validate
compile
test
package
verify
install
site
deploy
Plugins
Dependencies
Activer Windows
Accédez aux paramètres pour activer Windows.
9: Git
TODO
6: Problems
Terminal
Build
Event Log
master

On the bottom of the window, you will see the building process, wait for an `[INFO] BUILD SUCCESS` message.



```
[INFO] META-INF/native/ already added, skipping
[INFO] org/ already added, skipping
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time:  18.282 s
[INFO] Finished at: 2020-10-20T16:18:12+02:00
[INFO] ------------------------------------------------------------------------
```

You will see a new folder `package`, inside this folder you will find the JAR. We will use the JAR `*-with-dependencies.jar`.



```
[INFO] META-INF/native/ already added, skipping
[INFO] org/ already added, skipping
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time:  18.282 s
[INFO] Finished at: 2020-10-20T16:18:12+02:00
[INFO] ------------------------------------------------------------------------
```

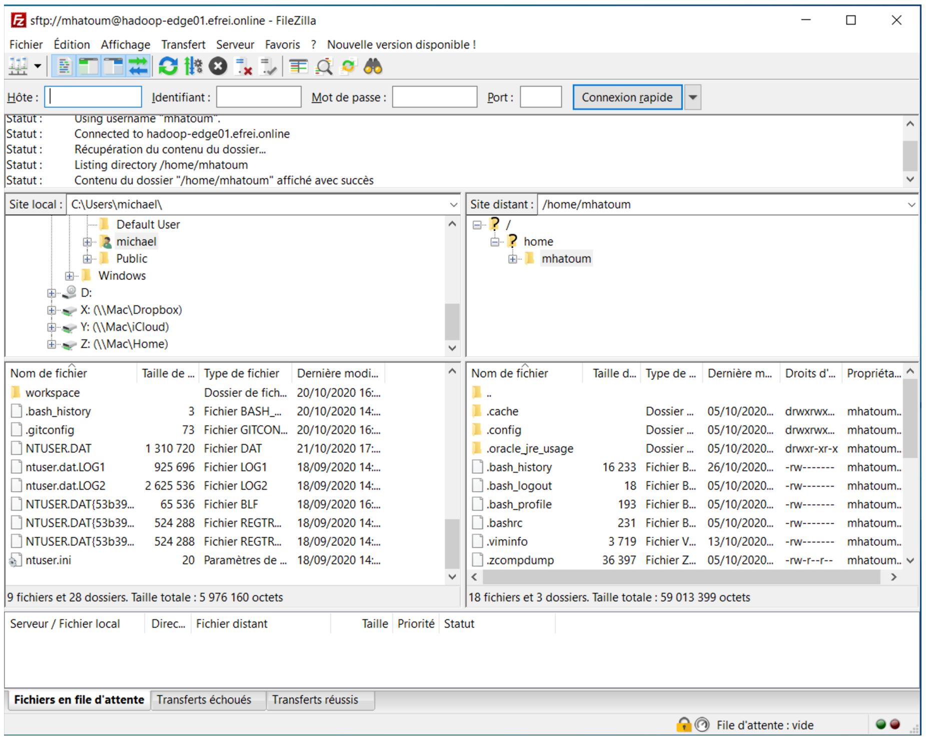
1.6 Send the JAR to the edge node

1.6.1 Microsoft Windows

Download and install FileZilla client for Windows from here. Connect to the edge node using thoses parameters:

- *Host*: `sftp://hadoop-edge01.efrei.online`
- *Username*: `ssh username`
- *Password*: `ssh password`
- *Port*: 22

You can slide files from left (your directories) to the right (edge node).


sftp://mhatoum@hadoop-edge01.efrei.online - FileZilla
Fichier Édition Affichage Transfert Serveur Favoris ? Nouvelle version disponible !
Hôte : Identifiant : Mot de passe : Port : Connexion rapide
Site local : C:\Users\michael\
Site distant : /home/mhatoum
Fichiers en file d'attente
Transferts échoués
Transferts réussis
File d'attente : vide

```
Statut :      Using username "mhatoum".
Statut :      Connected to hadoop-edge01.efrei.online
Statut :      Récupération du contenu du dossier...
Statut :      Listing directory /home/mhatoum
Statut :      Contenu du dossier "/home/mhatoum" affiché avec succès
```

```
    Default User
    michael
    Public
  Windows
D:
X: (\\Mac\Dropbox)
Y: (\\Mac\iCloud)
Z: (\\Mac\Home)
```

```
/
  home
    mhatoum
```

Nom de fichier	Taille de ...	Type de fichier	Dernière modi...
workspace		Dossier de fich...	20/10/2020 16:...
.bash_history	3	Fichier BASH_...	20/10/2020 14:...
.gitconfig	73	Fichier GITCON...	20/10/2020 16:...
NTUSER.DAT	1 310 720	Fichier DAT	21/10/2020 17:...
ntuser.dat.LOG1	925 696	Fichier LOG1	18/09/2020 14:...
ntuser.dat.LOG2	2 625 536	Fichier LOG2	18/09/2020 14:...
NTUSER.DAT{53b39...	65 536	Fichier BLF	18/09/2020 16:...
NTUSER.DAT{53b39...	524 288	Fichier REGTR...	18/09/2020 14:...
NTUSER.DAT{53b39...	524 288	Fichier REGTR...	18/09/2020 14:...
ntuser.ini	20	Paramètres de ...	18/09/2020 14:...

9 fichiers et 28 dossiers. Taille totale : 5 976 160 octets

Nom de fichier	Taille d...	Type de ...	Dernière m...	Droits d'...	Propriéta...
..					
.cache		Dossier ...	05/10/2020...	drwxrwx...	mhatoum...
.config		Dossier ...	05/10/2020...	drwxrwx...	mhatoum...
.oracle_jre_usage		Dossier ...	05/10/2020...	drwxr-xr-x	mhatoum...
.bash_history	16 233	Fichier B...	26/10/2020...	-rw-------	mhatoum...
.bash_logout	18	Fichier B...	05/10/2020...	-rw-------	mhatoum...
.bash_profile	193	Fichier B...	05/10/2020...	-rw-------	mhatoum...
.bashrc	231	Fichier B...	05/10/2020...	-rw-------	mhatoum...
.viminfo	3 719	Fichier V...	13/10/2020...	-rw-------	mhatoum...
.zcompdump	36 397	Fichier Z...	05/10/2020...	-rw-r--r--	mhatoum...

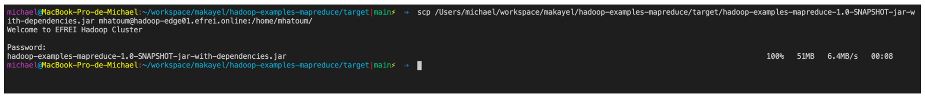
18 fichiers et 3 dossiers. Taille totale : 59 013 399 octets

Serveur / Fichier local	Direc...	Fichier distant	Taille	Priorité	Statut

1.6.2 Mac OS X

On Unix based operating system, you can use the `scp` command. Here an example:

```
$ scp myjar.jar username@hadoop-edge01.efrei.online:/home/username/
```



```
michael@MacBook-Pro-de-Michael:~/workspace/makayel/hadoop-examples-mapreduce/target|main⚡ ⇒ scp /Users/michael/workspace/makayel/hadoop-examples-mapreduce/target/hadoop-examples-mapreduce-1.0-SNAPSHOT-jar-w
ith-dependencies.jar mhatoum@hadoop-edge01.efrei.online:/home/mhatoum/
Welcome to EFREI Hadoop Cluster

Password:
hadoop-examples-mapreduce-1.0-SNAPSHOT-jar-with-dependencies.jar                                                   100%   51MB   6.4MB/s   00:08
michael@MacBook-Pro-de-Michael:~/workspace/makayel/hadoop-examples-mapreduce/target|main⚡ ⇒
```

1.6.3 Run the job

Start the `wordcount` job using the JAR that we send to the edge node. Referer to previous labs for the command.

1.7 How will the lab work?

1.7.1 Important 1

This lab will last at least 2 weeks and there are all kinds of exercises, easy and hard. So start with all the easy exercises, you will do the hard ones next time.

If you start with the difficult ones and you have no experience, you are going to have problems. *You do as you want, it's up to you.*

For each of these projects, if you can't do everything that is asked for, try to get closer. The scoring takes into account your attempts.

1.7.2 Important 2

Don't forget to write unit test if you want bonus points. Here some documentation about it.

1.7.3 Important 3

If you don't have a GitHub account, you must create one. You must clone the base repository following this tutorial (keep your repository public).

At the end of each lab session, you must add on Moodle a file with your commit ID:

```
$ <name1>-<name2>-<commit-id >.md
```

1.8 Remarkable trees of Paris

You are going to write some *MapReduce jobs* on the remarkable trees of Paris using this dataset.

Download the file and `put` it in your HDFS home directory.

Remember to ignore the first row in every mapper on this dataset.

1.8.1 Districts containing trees (very easy)

Write a MapReduce job that displays the list of distinct containing trees in this file. Obviously, it's twenty or less different arrondissements, but exactly how many?

You just need to put rounding as a key and put any value or NullWritable as a value, output from the mapper.

The reducer will just have to output the keys and values it receives, it doesn't even have to loop through the values since it ignores them.

1.8.2 Show all existing species (very easy)

Write a MapReduce job that displays the list of different species trees in this file.

1.8.3 Number of trees by kinds (easy)

Write a MapReduce job that calculates the number of trees of each kinds.

For example, there are 3 Acer, 19 Platanus, etc. How will you define the keys and values passed between the mapper and the reducer?

The mapper must extract the kind of tree. The reducer gets the pairs (`keyI`, `valueI`) with the same key, so this key must be the kind of tree; the value being the number of such trees.

1.8.4 Maximum height per kind of tree (average)

Write an MapReduce job that calculates the height of the tallest tree of each kind.

For example, the tallest Acer is 16m, the tallest Platanus is 45m, etc.

1.8.5 Sort the trees height from smallest to largest (average)

Write an MapReduce job that sort the trees height from smallest to largest.

1.8.6 District containing the oldest tree (difficult)

Write a MapReduce job that displays the district where the oldest tree is.

The mapper must extract the age and district of each tree.

The problem is, this information can't be used as keys and values (why?). You will need to define a subclass of Writable to contain both information.

The reducer should consolidate all this data and only output district.

1.8.7 District containing the most trees (very difficult)

Write a MapReduce job that displays the district that contains the most trees.

The problem is that the program will almost certainly display a list of pairs (district number, number of trees) not ordered by number.

If we apply this program to real big data not limited to the arrondissements of Paris, we would recover a huge list, unusable and the classification could take hours. How do you keep only the best answer?

The solution probably goes through a second MapReduce phase, in which the Mapper retrieves the pairs (district, number) from the first phase, and makes them pairs whose key is unimportant (NullWritable) and the values themselves are the same input pairs.

The Reducer receives them all and must keep the best pair.