**Setup TPs Spark MS BGD (2018-2019)**

**Setup TP 1: (MAC et Linux) Installation de Spark**

**I - Installation de java**

**Selon votre machine (mac, Linux perso, machine de TP) reportez-vous à la section correspondante dans la suite !**

* **Sur les machines de TP**

Dans un terminal, entrez (sans le “>” qui indique juste qu’on entre des commandes dans le terminal...) :

> java -version

Cette commande affiche la version de java qui est installée: la version 1.7 ou 1.8 doit déjà être installée.

* **Sur Ubuntu (machines perso : besoin des droits root)**

*Option 1:*

<http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html>

Dans “Java SE Development Kit 8u181” choisissez la version à dowloader qui vous correspond.

Puis installer le paquet downloadé

*Option 2:*

Dans un terminal:

> sudo add-apt-repository ppa:webupd8team/java  
> sudo apt-get update  
> sudo apt-get install oracle-java8-installer

* **Sur MAC**

<http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html>

Dans “Java SE Development Kit 8u181” télécharger le .dmg pour mac, puis l’installer.

Si la version de java n’est pas la bonne après installation (java -version). Dans le fichier .bash\_profile ajouter la ligne:

export JAVA\_HOME=`/usr/libexec/java\_home -v 1.8`

Puis fermer et réouvrir le terminal pour que la modification soit effective.

**II - Installation de Spark**

<http://spark.apache.org/downloads.html>

=> Spark release : 2.2.0

=> Package type : pre-build for hadoop 2.7

=> cliquer sur le lien : [spark-2.2.0-bin-hadoop2.7.tgz](https://archive.apache.org/dist/spark/spark-2.2.0/spark-2.2.0-bin-hadoop2.7.tgz)

Une fois téléchargé, copier le fichier “.tgz” dans votre répertoire “home” (dans un terminal entrez: echo $HOME pour savoir où est votre home). Puis décompresser le fichier “.tgz”, et c’est tout !

**Utiliser le Spark-shell**

Dans un terminal:

> cd spark-2.2.0-bin-hadoop2.7/bin

> ./spark-shell

L’interface Utilisateur est alors disponible dans un navigateur à l’adresse localhost:4040

**Réduire la quantité des logs affichés par Spark**

On copie le fichier de conf de logs par défaut “log4j.properties.template” dans “log4j.properties”.

Dans un terminal:

> cd spark-2.2.0-bin-hadoop2.7/conf

> cp log4j.properties.template log4j.properties

Dans un éditeur de texte ouvrez le fichier log4j.properties

Remplacez la ligne :

log4j.rootCategory=INFO, console

Par:

log4j.rootCategory=WARN, console

**Setup TP 2: Installation de SBT, IntelliJ et démarrage du projet**

Cette section est à faire pour le TP2 !

**I - Installation SBT et IntelliJ**

**Selon votre machine (mac, Linux perso, machine de TP) reportez-vous à la section correspondante dans la suite !**

* **Sur les machines de TP**

SBT est installé. Pour l’utiliser ouvrir un terminal et faire :

> sbt

IntelliJ 2016 nécessite java 1.8 (qui n’est pas installé sur les machines de TP).

Donc télécharger la version 15 de IntelliJ **Community** pour linux [ideaIC-15.0.6.tar.gz](https://download.jetbrains.com/idea/ideaIC-15.0.6.tar.gz) :

<https://confluence.jetbrains.com/display/IntelliJIDEA/Previous+IntelliJ+IDEA+Releases>

Décompresser IntelliJ

Dans un terminal:

> cd idea.../bin

> chmod +x idea.sh

> ./idea.sh

Dans la fenêtre qui s’ouvre:

=> I do not have previous installation => click OK

=> Choisir un thème => click next

=> create desktop entry : déselectionner “for all users” => click next

=> tune Idea to your task : ne rien faire => click next

=> scala : cliquer sur “Install” => start using IntelliJ

Une fois IntelliJ installé, aller dans les “préférences” puis dans “plugins” et dans la barre de recherche de plugins:

* Chercher “scala” et installer le plugins (si il ne l’est pas déjà)
* **Sur Ubuntu (machines perso : besoin des droits root)**

**Installation de sbt**

Dans un terminal:

> echo "deb https://dl.bintray.com/sbt/debian /" | sudo tee -a /etc/apt/sources.list.d/sbt.list  
> sudo apt-key adv --keyserver hkp://keyserver.ubuntu.com:80 --recv 2EE0EA64E40A89B84B2DF73499E82A75642AC823  
> sudo apt-get update  
> sudo apt-get install sbt

**Installation de IntelliJ Community edition**

Download :

<https://www.jetbrains.com/idea/>

Décompresser IntelliJ

Dans un terminal:

> cd idea-IC….etc…./bin

> ./idea.sh

Dans la fenêtre qui s’ouvre :

=> I do not have previous installation => click OK

=> Choisir un thème => click next

=> create desktop entry => click next

=> launcher script : ne rien faire => click next

=> tune Idea to your task : ne rien faire => click next

=> scala : cliquer sur “Install” => start using IntelliJ

Une fois IntelliJ installé, aller dans les “préférences” puis dans “plugins” et dans la barre de recherche de plugins:

* Chercher “scala” et installer le plugins (si il ne l’est pas déjà)
* **Sur MAC**

**Installer brew si ce n’est pas déjà fait:**

<https://brew.sh/index_fr>

**Installation de SBT**

Dans un terminal:

> brew install sbt

**Installation de IntelliJ community edition**

<https://www.jetbrains.com/idea/>

Télécharger le fichier .dmg, l’installer

Lancer IntelliJ, dans la fenêtre qui s’ouvre faire:

=> I do not have previous installation => click OK

=> Choisir un thème => click next

=> create desktop entry => click next

=> laucher script : ne rien faire => click next

=> tune Idea to your task : ne rien faire => click next

=> scala : cliquer sur “Install” => start using ItelliJ

Une fois IntelliJ installé, aller dans les “préférences” puis dans “plugins” et dans la barre de recherche de plugins:

* Chercher “scala” et installer le plugins (si il ne l’est pas déjà)

**II - (Mac et Linux) Importer le projet (voir TP 2 pour télécharger le template de projet) dans IntelliJ**

Ouvrir IntelliJ

=> Import project

=> Import project from external model, et choisir SBT

=> sélectionner tp\_spark/tp\_spark

=> sélectionner “use auto import” / project SDK cliquer sur “new” puis “JDK” sélectionner “java-8-oracle” dans l’arborescence / cliquer sur Finish.

=> sbt data project to import, ne rien faire, cliquer sur OK

.

. Attendre

.

**HOW TO: lancer un job Spark**

**Compiler et construire le jar**

Dans un terminal :

> cd tp\_spark/tp\_spark (aller là où se trouve le fichier build.sbt du projet)

> sbt assembly

L’adresse du jar est donnée vers la fin du script :

[info] Packaging /home/max/tp\_spark/tp\_spark/target/scala-2.11/tp\_spark-assembly-1.0.jar

**Démarer un cluster Spark local (le driver et le worker seront sur la même machine)**

Dans un terminal:

> cd spark-2.2.0-bin-hadoop2.7/sbin (attention c’est bien **“sbin”**)

> ./start-all.sh

[ Si il y a une erreur “port 22 connection refused”, c’est que le worker ne trouve pas l’adresse du master, ils ne peuvent donc pas communiquer. Pour démarrer le cluster il faut alors faire (toujours dans un terminal, et dans le dossier sbin): > ./start-master.sh ]

Allez à l’adresse localhost:8080 dans un navigateur, repérez l’adresse en gras tout en haut (spark://adresse\_du\_master:7077). Notez qu’il n’y a pas de worker indiqué sous worker Id. Puis retournez dans le terminal et faîtes:

> ./start-slave.sh adresse\_du\_master:7077

Il devrait maintenant y avoir un worker indiqué sous worker Id !

Dans chrome, firefox, etc:

Aller à l’adresse localhost:8080

L’Interface Utilisateur (Spark UI) s’affiche si spark a bien démarré.

**Soumettre un Job à Spark**

Soumettre le jar du script qui a été compilé:

Dans un terminal:

> cd spark-2.2.0-bin-hadoop2.7/bin ( !!!! Attention c’est bien “bin” maintenant)

> ./spark-submit --driver-memory 3G --executor-memory 4G --class com.sparkProject.Job --master spark://MBP-de-maxime-2:7077 /Users/maxime/IdeaProjects/tp\_spark/target/scala-2.10/tp\_spark-assembly-1.0.jar

(Remplacer “MBP-de-maxime” par ce qui est affiché tout en haut de votre Spark UI, remplacer le chemin vers le jar, remplacer le nom de la classe si il a été modifié par rapport au template donné en début de TP)

> ./spark-submit --conf spark.eventLog.enabled=true --conf spark.eventLog.dir="/tmp" --driver-memory 3G --executor-memory 4G --class com.sparkProject.Job --num-executors 2 --packages "com.amazonaws:aws-java-sdk:1.7.4,org.apache.hadoop:hadoop-aws:2.7.1" --master spark://MBP-de-maxime-2:7077 /Users/maxime/IdeaProjects/tp\_spark/target/scala-2.10/tp\_spark-assembly-1.0.jar