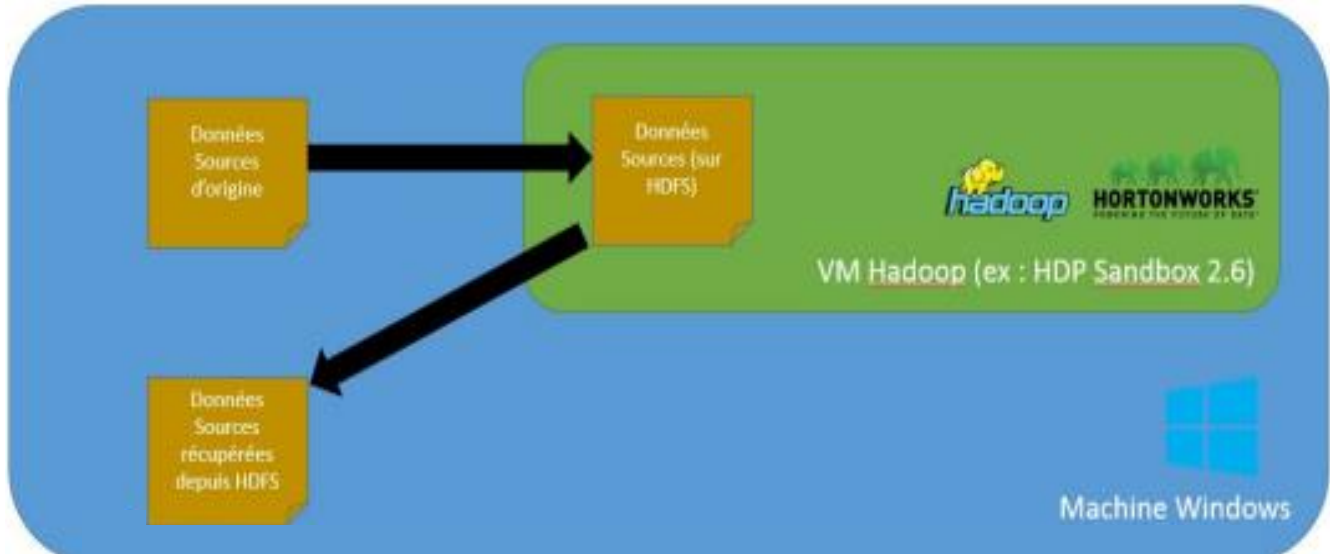


## Machine Windows

### Préparation et configuration de l'environnement



#### Installation de la VM Hadoop :

Pour la mise en place de la VM Hadoop, nous allons utiliser Hortonworks Sandbox. Pour cela téléchargez Hortonworks HDP sur le lien suivant :

<https://www.cloudera.com/downloads/hortonworks-sandbox.html>

Ensuite téléchargez VirtualBox, le téléchargement se lancera automatiquement avec ce lien :

<https://download.virtualbox.org/virtualbox/6.1.22/VirtualBox-6.1.22-144080-Win.exe>

Pour voir toutes les étapes pour l'installation de la VM Hadoop sur VirtualBox, ouvrez le document *instr\_hadoop\_vb.pdf* fournit.

Vous devez avoir à partir de ce point une VM Hadoop fonctionnelle.

#### Mise en place des fichiers sur la machine Windows :

Le but de cette partie est de pouvoir échanger des fichiers entre la machine Windows et HDFS dans la VM Hadoop.

Pour cela, nous utilisons 2 fichiers :

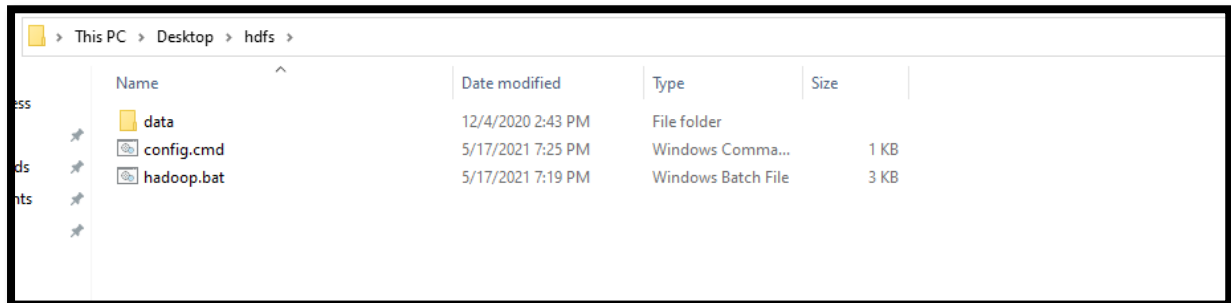
- Hadoop.bat, qui contient les commandes permettant de faire le transfert des fichiers entre la machine Windows et HDFS.
- Config.cmd, qui contient les variables pour le bon fonctionnement de hadoop.bat.

Pour les mettre en place, mettez-les dans un dossier où vous le souhaitez.

Placez également dans ce dossier le dossier data contenant les 3 fichiers de données : *data.json*, *label.csv* et *categories\_string.csv*.

Pour le dossier data, il est conseillé de le mettre au même niveau que les autres fichiers mais vous pouvez différemment (le chemin vers le dossier data est paramétrable).

Ce qui nous donne ceci :



Ouvrez ensuite le fichier *config.cmd* et modifier les valeurs des variables pour les faire correspondre à votre cas :

- La variable « *local\_default\_directory* » correspond au dossier par défaut. Celui-ci est sous la forme *C:\Users\nom\_utilisateur*. Pour l'exemple dans notre cas c'est : *C:\Users\revin*
- La variable « *path\_files\_to\_transfer\_directory* » correspond au dossier qui contient les fichiers à transférer. Si vous avez suivi ce guide, il est sous la forme *C:\Users\nom\_utilisateur\big\_data\_project\data*. Pour exemple dans notre cas c'est : *C:\Users\revin\big\_data\_project\data*
- Les variables liées aux identifiants « *username* » et « *password* » (de base *maria\_dev*) :
- La variable « *folder\_name\_store\_data* » correspond au nom du dossier qui contiendra tous les fichiers de données dans chaque environnement.
- La variable « *path\_hdfs* » qui correspond au chemin dans HDFS où entreposer les données.

La mise en place des fichiers est ainsi terminée.

## Echange de fichiers entre la machine Windows et HDFS (VM Hadoop)

A ce point dans le guide d'installation, vous devez avoir une VM Hadoop lancée et fonctionnelle, ainsi que vos fichiers mis en place sur la machine Windows.

Il suffit maintenant d'ouvrir une invite de commande sous Windows 10 et de lancer cette ligne de commande : `start big_data_project/hadoop.bat`

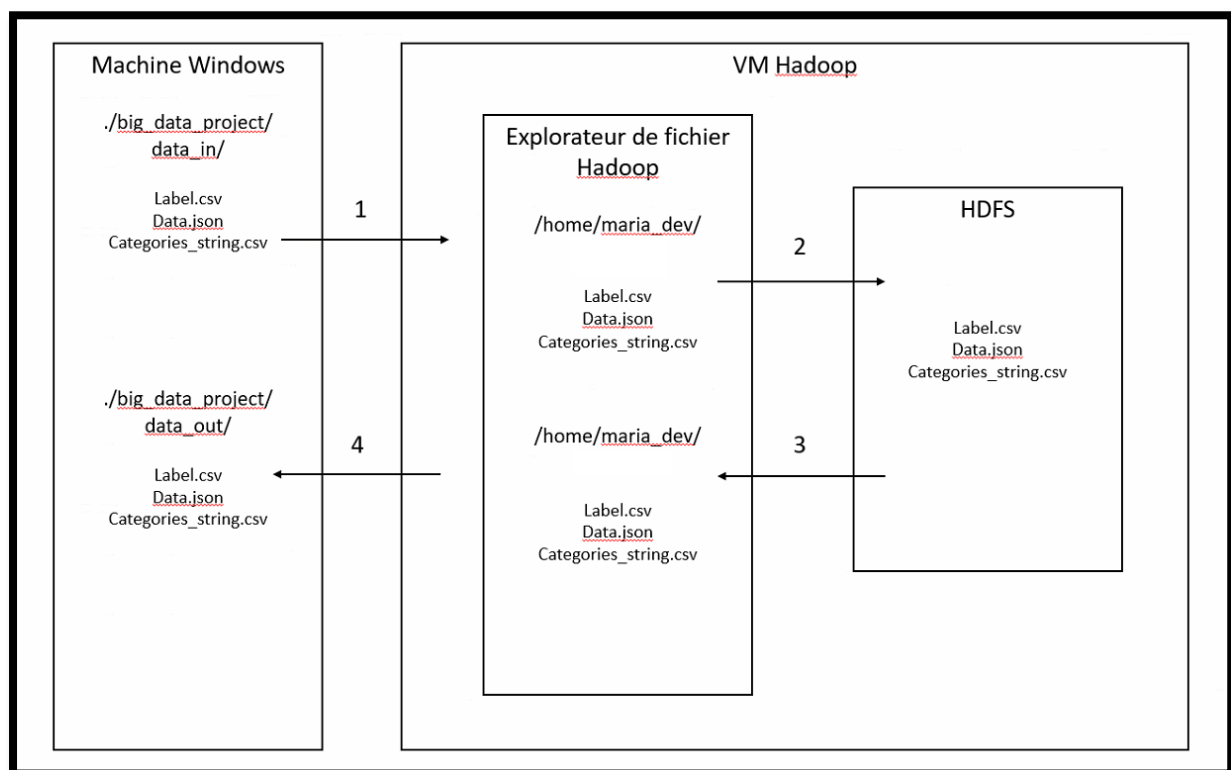
**Important** : Lors de l'exécution vous allez devoir effectuer quelques tâches :

- La première sera de mettre Yes or No, mettez Yes
- La deuxième, certaines commandes demandent un mot de passe, ce dernier est : `maria_dev`

### **Fonctionnement et architecture finale:**

Vous aurez à la fin du script dans `C:/Users/nom_utilisateur/big_data_project/data_in/` les documents qui ont été envoyés et dans `C:/Users/nom_utilisateur/big_data_project/data_out/` les documents qui ont été reçus.

Voici un schéma qui reprend les étapes et les différents chemins où se trouvent les documents dans les différents environnements :



## Quelques informations supplémentaires

### Accéder à la VM Hadoop :

Pour accéder à la VM Hadoop, il suffit de lancer un navigateur internet (chrome par exemple) et d'accéder à cette URL : <http://localhost:4200>

Une fois sur cette nouvelle page, vous allez devoir vous connecter pour cela utiliser ces identifiants :

Login : maria\_dev

Mot de passe : maria\_dev

Vous saurez ainsi connecter à la VM Hadoop et pourrez naviguer dans le système d'exploration de fichiers. Vous pouvez par exemple visualiser un des fichiers que nous avons transférés plus tôt dans ce guide. Pour cela tapez cette commande : `tail -n50 /home/maria_dev/data_in/label.csv` qui permet d'afficher les 50 dernières lignes du fichier *label.csv*.

### Accéder au Dashboard de la VM Hadoop :

Pour accéder au Dashboard de la VM Hadoop, il suffit de lancer un navigateur internet (chrome par exemple) et d'accéder à cette URL : <http://localhost:1080>

Une fois sur cette nouvelle page, vous allez devoir vous connecter pour cela utiliser ces identifiants :

Login : maria\_dev

Mot de passe : maria\_dev

Vous aurez ainsi accès aux informations de la VM Hadoop.

### Accéder à HDFS :

Pour accéder à HDFS et consulter un document par exemple, il faut d'abord se connecter à la VM Hadoop (voir Accéder à la VM Hadoop).

Ensuite taper la ligne suivante : `hadoop fs -ls` qui permet de voir la liste des fichiers et dossiers présents dans HDFS.

# Procédure de mise en place de la partie Cloud

Voici les différentes étapes :

- Ouvrez variables.tf et modifiez les valeurs avec vos informations de connexion
- Ouvrez ensuite le fichier « config\_cloud.cmd » et modifiez le chemin vers la clé pem.  
Exemple: C:\Users\thiba\big\_data\_project\Connexion\rvt-key-pair.pem
- Double-cliquez sur le fichier terraform.bat

Une connexion SSH va s'ouvrir vers la VM via Putty.

Ensuite une fois sur la VM :

Il faut utiliser les commandes suivantes pour configurer la VM :

```
sudo amazon-linux-extras enable python3.8
```

```
sudo yum install python3.8
```

```
sudo update-alternatives --install /usr/bin/python python  
/usr/bin/python3.8 1
```

```
sudo update-alternatives --list | grep python
```

```
sudo curl https://bootstrap.pypa.io/get-pip.py -o get-pip.py  
sudo python get-pip.py
```

```
pip install pandas
```

```
pip install nltk
```

```
pip install sklearn
```

```
pip install counter
```

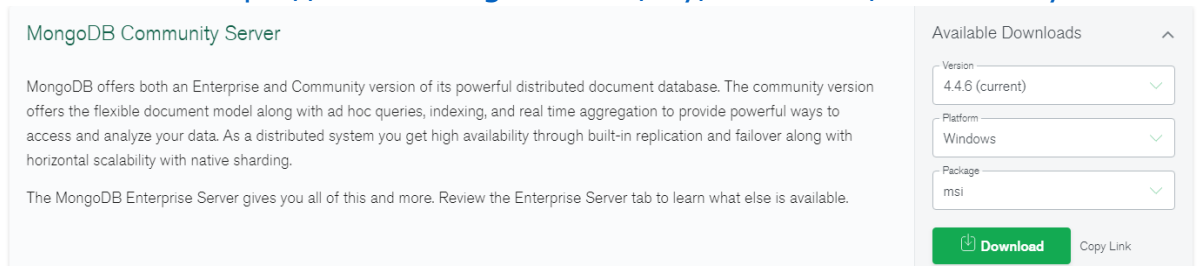
Vous pouvez maintenant lancer le script Python n'importe où sur l'instance EC2. Il se trouve dans le dossier SCRIPT\_PYTHON.

# Procédure de mise en place de la partie MongoDB

Voici les différentes étapes :

- Téléchargez AWS CLI. Vous pouvez le faire à l'aide de cette commande : `msiexec.exe /i https://awscli.amazonaws.com/AWSCLIV2.msi`

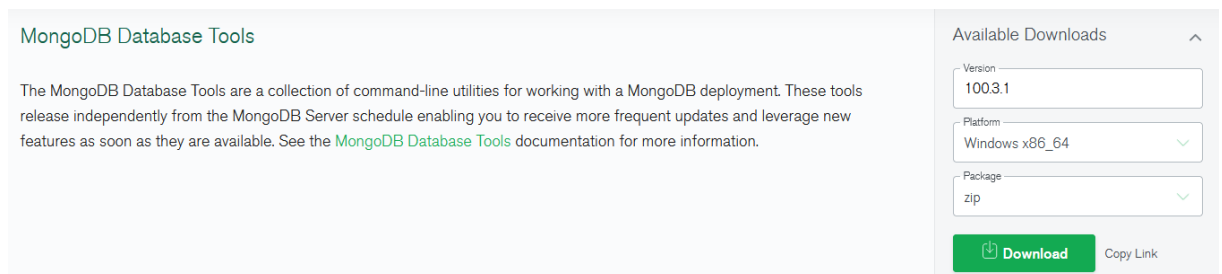
- Allez sur le site : <https://www.mongodb.com/try/download/community>



The screenshot shows the 'MongoDB Community Server' download page. On the left, there is a description of the community version. On the right, under 'Available Downloads', there are three dropdown menus: 'Version' set to '4.4.6 (current)', 'Platform' set to 'Windows', and 'Package' set to 'msi'. Below these is a green 'Download' button and a 'Copy Link' link.

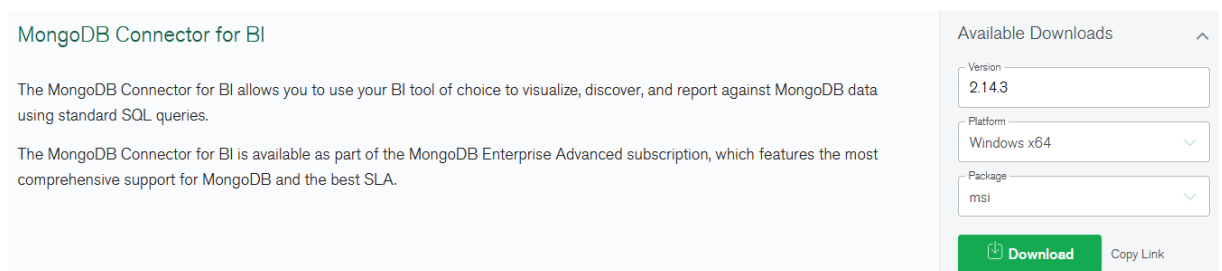
- Cliquez sur « Download »

- Allez ensuite sur le site : <https://www.mongodb.com/try/download/database-tools>



The screenshot shows the 'MongoDB Database Tools' download page. On the right, under 'Available Downloads', there are three dropdown menus: 'Version' set to '100.3.1', 'Platform' set to 'Windows x86\_64', and 'Package' set to 'zip'. Below these is a green 'Download' button and a 'Copy Link' link.

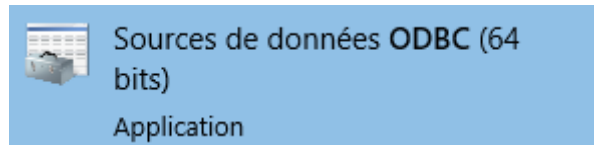
- Allez ensuite sur le site : <https://www.mongodb.com/try/download/bi-connector>



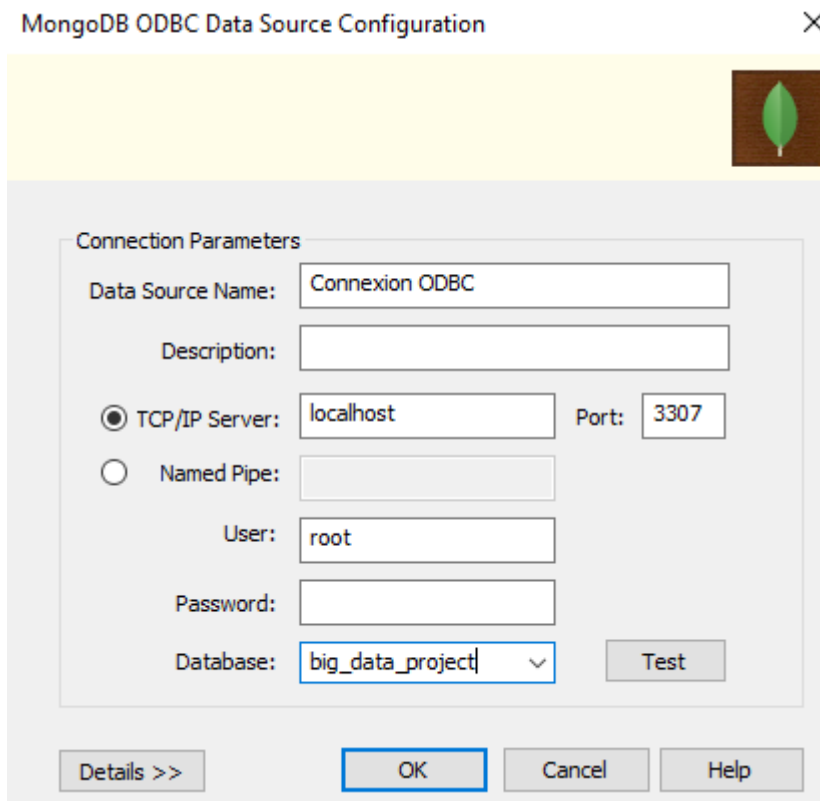
The screenshot shows the 'MongoDB Connector for BI' download page. On the right, under 'Available Downloads', there are three dropdown menus: 'Version' set to '2.14.3', 'Platform' set to 'Windows x64', and 'Package' set to 'msi'. Below these is a green 'Download' button and a 'Copy Link' link.

- Cliquez sur « Download »
- Allez ensuite sur ce lien : <https://github.com/mongodb/mongo-odbc-driver/releases>
- Téléchargez `mongodb-connector-odbc-1.4.2-win-64-bit.msi`


- Vous devez ensuite mettre en place une nouvelle connexion sur ODBC.



- Voici les informations que vous devez mettre :



- Posez maintenant le fichier « recuperation\_cloud\_et\_creation\_mongo.bat » dans le dossier : C:\Program Files\MongoDB\Server\4.4\bin
- Vous devez aller dans C:\Users\User\.aws et modifier le fichier « credentials » de la façon suivante :

 credentials - Bloc-notes

Fichier Edition Format Affichage Aide

[default]

```
aws_access_key_id = PUT_YOUR_ACCESS_KEY_ID_HERE
aws_secret_access_key = PUT_YOUR_SECRET_KEY_ID_HERE
aws_session_token= PUT_YOUR_SESSION_TOKEN_HERE
```

- Double-cliquez sur le fichier « recuperation\_cloud.bat ». Vous devriez voir les fichiers de résultats apparaître dans le dossier. Ils ont été récupérés depuis le Cloud. Ils ont aussi été ajoutés sur MongoDB sous forme de collections.
- Allez dans le dossier : « C:\Program Files\MongoDB\Connector for BI\2.14\bin » et double-cliquez sur « mongosql.exe ». Vous pourrez maintenant vous connecter à la base MongoDB avec Power BI.



# Installation et utilisation de Power BI

## Installation du logiciel

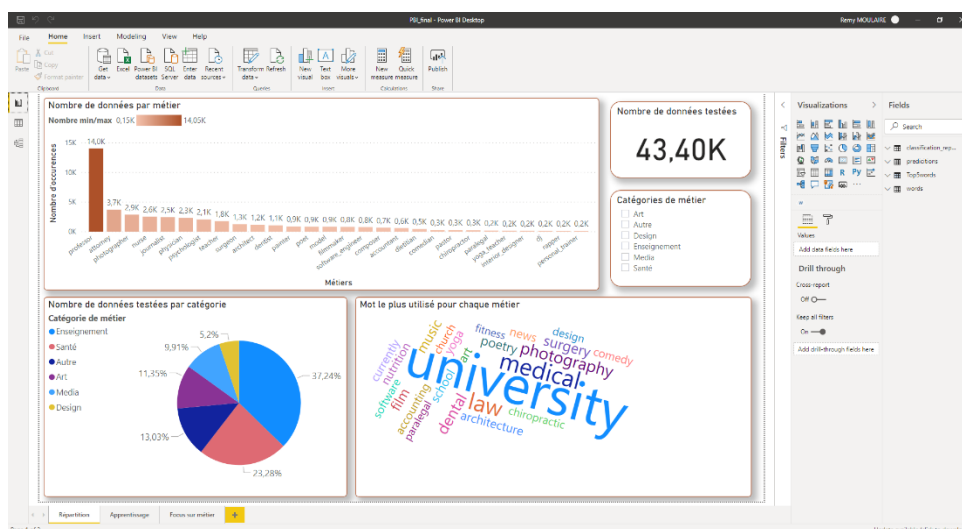
Assurez-vous avant de suivre cette procédure que vous êtes bien sur votre machine Windows sur laquelle vous avez implémenté la base MongoDB.

- Téléchargez l'outil Power BI au lien suivant : <https://www.microsoft.com/en-us/download/confirmation.aspx?id=58494>
- Suivez le flux d'installation jusqu'à ce que le logiciel soit installé sur votre machine.
- Lancez Power BI.

## Utilisation du logiciel

- Une fois sur l'outil, cliquez sur « Fichier », « Ouvrir un rapport ».
- Sélectionnez le fichier .pbix fourni avec cette procédure dans le dossier « DATAVISUALISATION ».
- Lorsque le rapport est ouvert, les rapports devraient se mettre à jour avec les données de votre base MongoDB, la connexion étant déjà configurée.

Voici à quoi doit ressembler le rapport :



Pour manipuler les données en entrée ou configurer un autre point de connexion, utilisez le bouton « transformer les données » sur le panel du haut.

Pour tout autre usage, la documentation sur l'outil Power BI est trouvable sur le lien suivant :

<https://docs.microsoft.com/fr-fr/power-bi/fundamentals/desktop-getting-started>