



Premiers pas sur la plateforme Data Engineer

🕒 30 minutes 📺 Easy



DataScientest • com

Premiers pas sur la plateforme Data Engineer

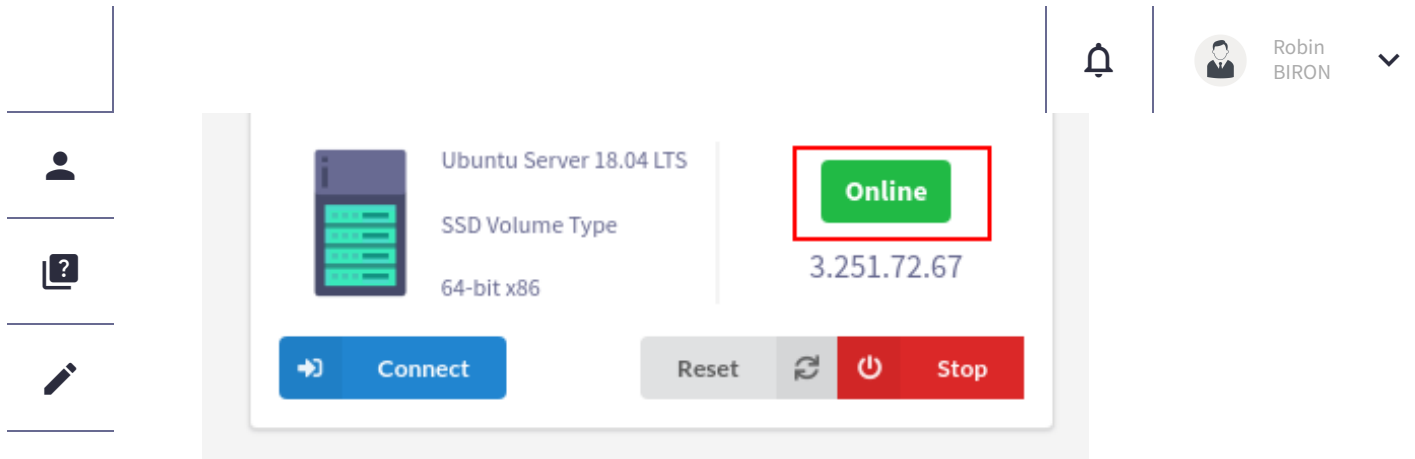
L'utilisation de certains outils ne peut se démontrer en utilisant des notebooks Jupyter. Dans le cadre des formations Data Engineering/Machine Learning Engineer, nous avons choisi de mettre à disposition des machines virtuelles hébergées dans le cloud. Ces machines comportent les outils nécessaires au déroulement des cours.

On peut choisir de suivre ces cours sur ces machines virtuelles distantes ou directement sur sa machine locale. Il faudra simplement installer les outils suivants:

- Python3, PIP, VirtualEnv, IPython
- Docker, Docker-Compose

État de la machine virtuelle

L'état de la machine virtuelle est indiqué dans l'encadré "Machine Status" en haut à gauche de la page de cours :



On peut ainsi voir si elle est arrêtée ou en fonctionnement.

On distingue également 3 boutons:

- **Connect**: pour afficher des informations sur la connexion à la machine distante.
- **Reset**: pour réinitialiser la machine à la fin d'un cours. Attention, cette opération supprime tous les fichiers que vous avez ajoutés sur la machine.
- **Stop**: pour arrêter la machine lorsqu'on ne l'utilise pas.

Se connecter à la machine distante

Pour se connecter à la machine virtuelle, on va utiliser une connexion SSH. La connexion SSH permet de créer un canal de communication sécurisé entre deux machines: une machine locale, c'est-à-dire l'ordinateur sur lequel vous travaillez et une machine distante, le serveur auquel vous voulez vous connecter.

En cliquant sur le bouton **Connect**, on peut voir les instructions pour se connecter à la machine. L'outil le plus simple pour se connecter à ces machines est OpenSSH, installé par défaut sur les systèmes Linux et MacOS. À noter que depuis Windows 10, OpenSSH est installé par défaut également.

Robin
BIRON

Machine status



Ubuntu
Server
18.04
LTS
SSD
Volume
Type
64-bit
x86

Online

34.245.135.23



Connect

Reset



Stop



Ubuntu Server 18.04 LTS

SSD Volume Type

64-bit x86

Online

3.251.72.67



Connect

Reset



Stop

1. Download your private key file



data_engineering_machine.pem

Download

2. Your key must not be publicly viewable for SSH to work. Use this command if needed:

```
chmod 400 data_engineering_machine.pem
```

3. Connect to your instance

```
ssh -i "data_engineering_machine.pem"  
ubuntu@3.251.72.67
```

4. If you need to forward a process running on the distant machine, you can use **-L** argument

```
ssh -i data_engineering_machine.pem -L  
1234:localhost:4321 -L 5678:localhost:8765  
ubuntu@3.251.72.67
```

Distant port **1234** is forwarded to **localhost:4321**, **5678** to **localhost:8765**

Téléchargement de la clef privée

Pour authentifier la connexion, nous allons utiliser une clef privée: elle est téléchargeable via le bouton **Download** ou en utilisant ce lien.

macOs / Linux

Une fois téléchargée, nous devons changer les droits sur cette clef. Cette opération n'est à faire **qu'une seule fois** et se réalise depuis un terminal. Il nous faut donc ouvrir un terminal et se déplacer jusque dans le dossier contenant la clef.

Sur Linux, le terminal s'ouvre en utilisant la commande **ctrl + alt + T**. Sur macOS vous pourrez l'ouvrir depuis le Launchpad.

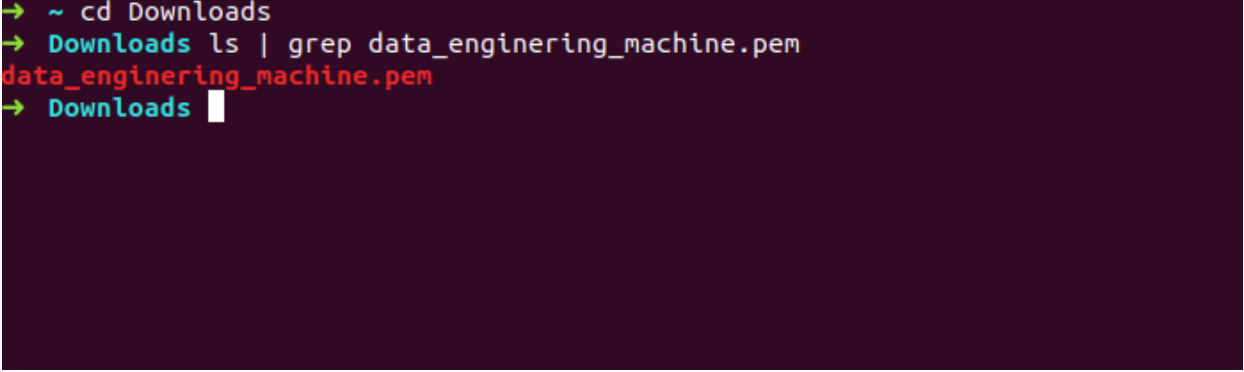
On utilisera la commande **cd** pour se déplacer dans les dossiers. Par exemple, on pourra utiliser la commande suivante pour se déplacer dans le dossier qui contient la clef privée que vous venez de télécharger (par exemple **/home/username/Downloads**):

```
1 cd /home/username/Downloads
```



utilisant la commande suivante:

```
1 ls | grep data_engineering_machine.pem
2
```



```
→ ~ cd Downloads
→ Downloads ls | grep data_engineering_machine.pem
data_engineering_machine.pem
→ Downloads
```

Si cette commande affiche un résultat alors la clef est bien présente dans le dossier.

On va ensuite changer les droits en lecture de la clef en utilisant la commande suivante:

```
1 chmod 400 data_engineering_machine.pem
2
```

Windows

Une fois la clef téléchargée, vous devrez ouvrir un terminal en cherchant `cmd` dans le menu **Démarrer**. On pourra aussi utiliser la commande `cd` pour se déplacer dans les dossiers. Par exemple, pour se déplacer dans le dossier `C:\Users\Username\Downloads`, on pourra utiliser la commande:

```
1 cd C:\Users\Username\Downloads
2
```

Si vous souhaitez vérifier que la clef est bien présente, on pourra utiliser la commande:

```
1 dir | findstr data_engineering_machine.pem
2
```

Vous devrez également vérifier que le client OpenSSH est bien installé sur votre ordinateur, ce qui est le cas si vous utilisez une version récente de Windows.

Pour ce faire, vous pouvez directement taper `ssh` dans un terminal, et vous devriez obtenir une réponse similaire à celle-ci :



```
C:\Users\Fenton>ssh -i [key] [user@]host[:port]
usage: ssh [-46AaCfGgKkMnNqsTtVvXxYy] [-B bind_interface]
          [-b bind_address] [-c cipher_spec] [-D [bind_address:]port]
          [-E log_file] [-e escape_char] [-F configfile] [-I pkcs11]
          [-i identity_file] [-J [user@]host[:port]] [-L address]
          [-l login_name] [-m mac_spec] [-O ctl_cmd] [-o option] [-p port]
          [-Q query_option] [-R address] [-S ctl_path] [-W host:port]
          [-w local_tun[:remote_tun]] destination [command]
```

```
C:\Users\Fenton>_
```

Connexion à la machine

Une fois dans le dossier qui contient la clef privée, on peut se connecter en SSH à la machine virtuelle en utilisant la commande suivante:

```
1 ssh -i "data_engineering_machine.pem" ubuntu
2
```

On devra faire attention à bien préciser l'adresse IP de sa machine, affichée dans l'encadré "Machine Status". Ici, l'argument `-i` permet de spécifier le chemin vers le fichier de la clef privée. L'utilisateur qu'on choisit ici est `ubuntu`. Il n'y a pas de mot de passe pour se connecter avec cet utilisateur.

On pourra ouvrir plusieurs connexions en ouvrant différents terminaux.

Tunnels

Dans certains cours, on veut utiliser `ssh` pour faire suivre l'interface d'un processus tournant sur la machine distante vers la machine locale en utilisant l'argument `-L`. Par exemple, pour faire suivre un processus sur le port `1234` de la machine distante vers le port `4321` de la machine locale en utilisant la commande suivante:

```
1 ssh -i "data_engineering_machine.pem" -L 1234:localhost:4321
2
```

On pourra alors faire des requêtes à l'adresse locale `localhost:4321` qui seront redirigées via la connexion SSH vers le port `1234` de la machine distante.

Validated