

Installer Rstudio-server sur votre cluster et vous y connecter avec le terminal

Installer FoxyProxy

► Pour google chrome : [lien](#)

► Pour firefox: [lien](#)

Une fois FoxyProxy installé, ouvrez le plugin et importer le fichier se trouvant dans :

- <https://github.com/katossky/panorama-bigdata/blob/master/settings/foxyproxy-settings.json> pour firefox
- <https://github.com/katossky/panorama-bigdata/blob/master/settings/foxyproxy-settings.xml> pour chrome

Etablir une connexion SSH avec votre cluster

► La marche à suivre est également disponible si vous cliquez sur "SSH" depuis la page de votre cluster

► Vous pouvez vous connecter même si votre cluster est en "Démarrage en cours"

- ☐ Ouvrez un terminal
- ☐ Exécutez la commande suivante

```
ssh hadoop@[DNS public] -i ~/mykeypair.pem
```

Avec

- [DNS public] : le DNS public principal de votre cluster (vous le trouverez dans les informations de votre cluster sur l'interface aws). Vous pouvez noter le **[DNS public]** dans un bloc note pour le retrouver plus facilement.

The screenshot shows the Amazon EMR console interface. On the left is a navigation menu with options like 'Clusters', 'Configurations de la sécurité', 'Bloquer l'accès public', 'Sous-réseaux VPC', 'Événements', 'Aide', and 'Nouveautés'. The main content area shows details for a cluster named 'Mon cluster' which is in the 'En attente' (Waiting) state. At the top, there are buttons for 'Cloner', 'Réinitialiser', and 'Exporter AWS CLI'. Below these are tabs for 'Récapitulatif', 'Historique de l'application', 'Surveillance', 'Matériel', 'Configurations', 'Événements', 'Étapes', and 'Actions d'amorçage'. The 'Connexions' section shows a link to 'Activer la connexion Web'. The 'DNS public principal' is listed as 'ec2-35-180-228-218.eu-west-3.compute.amazonaws.com' with an 'SSH' icon next to it, which is highlighted by a red rectangle. Below this, there are sections for 'Récapitulatif' (ID, Date de création, Temps écoulé, Résiliation automatique, Protection de la désactivation) and 'Détails de configuration' (Étiquette de version, Distribution Hadoop, Applications, URI de connexion, Vue cohérente EMRFS, ID d'AMI personnalisée). On the right, the 'Réseau et matériel' section shows 'Zone de disponibilité', 'ID de sous-réseau (subnet)', 'Maître', 'Principal', and 'Tâche'.

- o ~/mykeypair.pem : le chemin vers la clef de votre cluster

riel	Sécurité
onibilité : eu-west-3b	Nom de clé : spark_cluster_TP
us-réseau subnet-638bf218 [?]	P
(subnet) :	Rôle EMR : EMR_DefaultRole
Maître : Action d'amorçage 1 m5.xlarge	Visible pour tous les Tous Modification
Principal : Mise en service 2 m5.xlarge	utilisateurs :
Tâche : -	Groupes de sécurité pour sg-049debc430865c513 [?]
	le principal : (ElasticMapReduce-master)
	Groupes de sécurité pour sg-081ef033d571d72c8 [?]
	la base et les tâches : (ElasticMapReduce-slave)

☐ Cliquez sur Yes pour ignorer l'alerte de sécurité.

Installer Rstudio server et postgresql-devel

```
sudo yum install libcurl-devel openssl-devel # used for devtools
```

```
hadoop@ip-172-31-9-22:~  
Installing:  
libcurl-devel      x86_64      7.61.1-12.93.amzn1      amzn-updates      855 k  
  
Transaction Summary  
=====  
Install 1 Package  
  
Total download size: 855 k  
Installed size: 1.3 M  
Is this ok [y/d/N]: y  
Downloading packages:  
libcurl-devel-7.61.1-12.93.amzn1.x86_64.rpm      | 855 kB    00:00  
Running transaction check  
Running transaction test  
Transaction test succeeded  
Running transaction  
  Installing : libcurl-devel-7.61.1-12.93.amzn1.x86_64      1/1  
  Verifying  : libcurl-devel-7.61.1-12.93.amzn1.x86_64      1/1  
  
Installed:  
libcurl-devel.x86_64 0:7.61.1-12.93.amzn1  
  
Complete!  
[hadoop@ip-172-31-9-22 ~]$
```

```
wget https://download2.rstudio.org/server/centos6/x86_64/rstudio-server-rhel-  
1.2.5033-x86_64.rpm  
sudo yum install rstudio-server-rhel-1.2.5033-x86_64.rpm
```

⚠ Faites bien attention, il est possible qu'en copiant/collant les lignes, un saut à la ligne se mette sur la première instruction, ce qui conduira à une erreur d'installation

```
hadoop@ip-172-31-9-22:~  
Transaction Summary  
-----  
Install 1 Package  
  
Total size: 306 M  
Installed size: 306 M  
Is this ok [y/d/N]: y  
Downloading packages:  
Running transaction check  
Running transaction test  
Transaction test succeeded  
Running transaction  
  Installing : rstudio-server-0.99.1266-1.x86_64 1/1  
groupadd: group 'rstudio-server' already exists  
rsession: no process found  
rstudio-server start/running, process 17084  
  Verifying : rstudio-server-0.99.1266-1.x86_64 1/1  
  
Installed:  
  rstudio-server.x86_64 0:0.99.1266-1  
  
Complete!  
[hadoop@ip-172-31-9-22 ~]$
```

- ☐ Créez un user pour Rstudio

```
# Make User  
sudo useradd -m rstudio-user  
sudo passwd rstudio-user
```

⚠ Retenez votre mot de passe

- ☐ Créez un dossier dans HDFS pour votre user

```
# Create new directory in hdfs  
hadoop fs -mkdir /user/rstudio-user  
hadoop fs -chmod 777 /user/rstudio-user
```

- ☐ Installez postgresql-devel

```
sudo yum install postgresql-devel
```

Etablir un tunnel SSH avec votre cluster

¶ La session SSH que vous allez ouvrir devra rester ouvertes tant que vous utiliser Rstudio-server

- ☐ Ouvrez un terminal
- ☐ Exécutez la commande suivante

```
ssh -i ~/mykeypair.pem -N -D 8157 hadoop@ec2-***[DNS public]**
```

Avec

- o [DNS public] : le DNS public principal de votre cluster (vous le trouverez dans les informations de votre cluster sur l'interface aws).
- o ~/mykeypair.pem : le chemin vers la clef de votre cluster

Normalement il ne va rien se passer, et vous pouvez croire que rien ne fonctionne. Mais il n'en est rien ! Si aucune erreur ne s'est affichée c'est que tout fonctionne

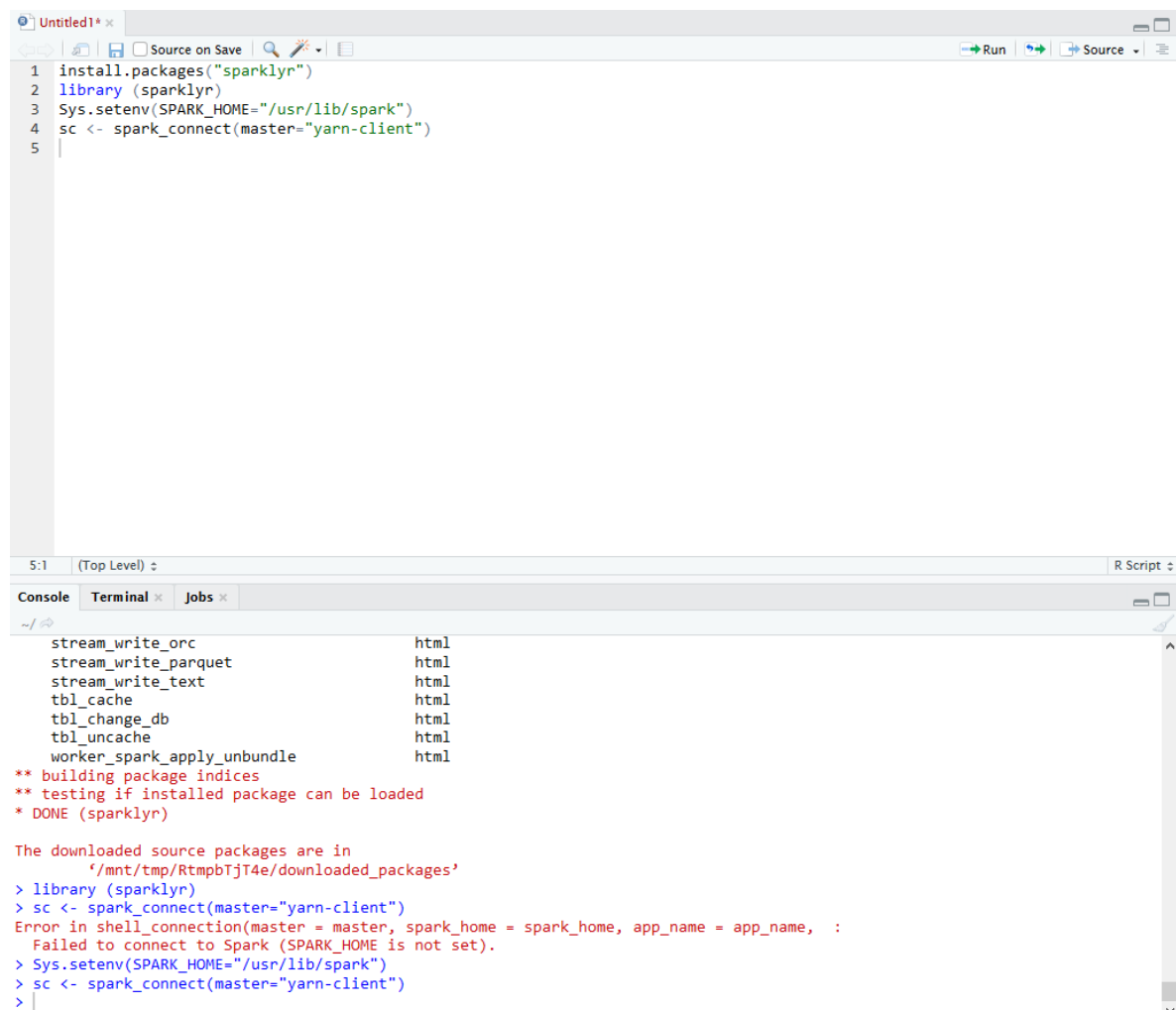
Se connecter à Rstudio-server

Connectez-vous à l'interface web de Rstudio server avec l'adresse suivante [https://\[DNS public\]:8787](https://[DNS public]:8787) avec [DNS public] le DNS public de votre cluster. Puis connectez vous avec l'utilisation rstudio-user et le mot de passe que vous avez choisi.

Se connecter au cluster spark via Rstudio-server

Voici un code minimal pour vous connecter au cluster spark avec Rstudio-server

```
install.packages("sparklyr")
library(sparklyr)
Sys.setenv(SPARK_HOME="/usr/lib/spark")
sc <- spark_connect(master="yarn-client")
```



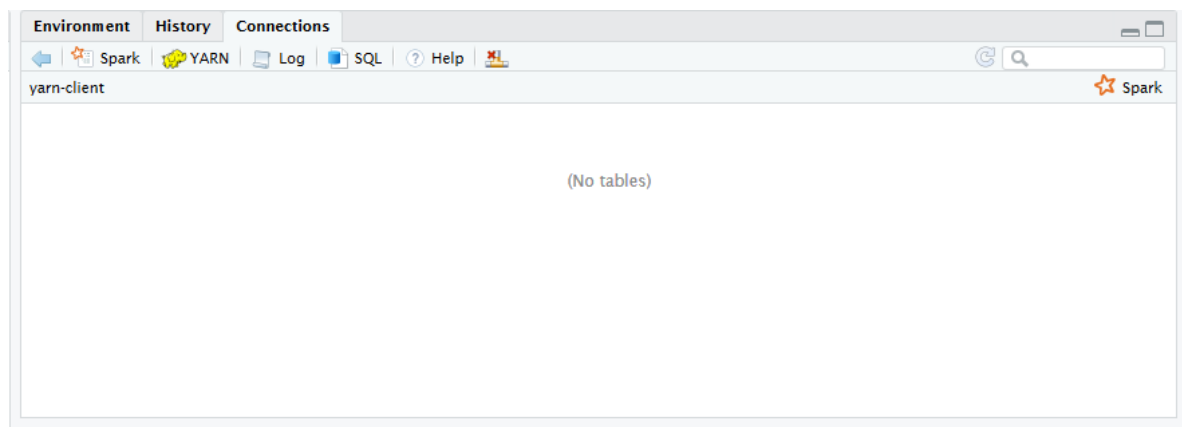
The screenshot shows the RStudio interface. The top pane is the script editor, titled 'Untitled1', containing the following R code:

```
1 install.packages("sparklyr")
2 library(sparklyr)
3 Sys.setenv(SPARK_HOME="/usr/lib/spark")
4 sc <- spark_connect(master="yarn-client")
5
```

The bottom pane is the console window, showing the output of the code execution. It displays a list of installed packages and their versions, followed by a message indicating that the Spark package is being tested and loaded successfully. The console output is as follows:

```
stream_write_orc          html
stream_write_parquet      html
stream_write_text         html
tbl_cache                 html
tbl_change_db             html
tbl_uncache               html
worker_spark_apply_unbundle html
** building package indices
** testing if installed package can be loaded
* DONE (sparklyr)

The downloaded source packages are in
  '/mnt/tmp/RtmpbTjT4e/downloaded_packages'
> library(sparklyr)
> sc <- spark_connect(master="yarn-client")
Error in shell_connection(master = master, spark_home = spark_home, app_name = app_name, :
  Failed to connect to Spark (SPARK_HOME is not set).
> Sys.setenv(SPARK_HOME="/usr/lib/spark")
> sc <- spark_connect(master="yarn-client")
>
```



🔔🔔🔔 Pensez à sauvegarder votre script sur votre poste de temps en temps via des copier coller.
Surtout avant d'éteindre votre cluster !

🔔🔔🔔🔔🔔🔔🔔🔔 **PENSEZ À ÉTEINDRE VOTRE CLUSTER À LA FIN**