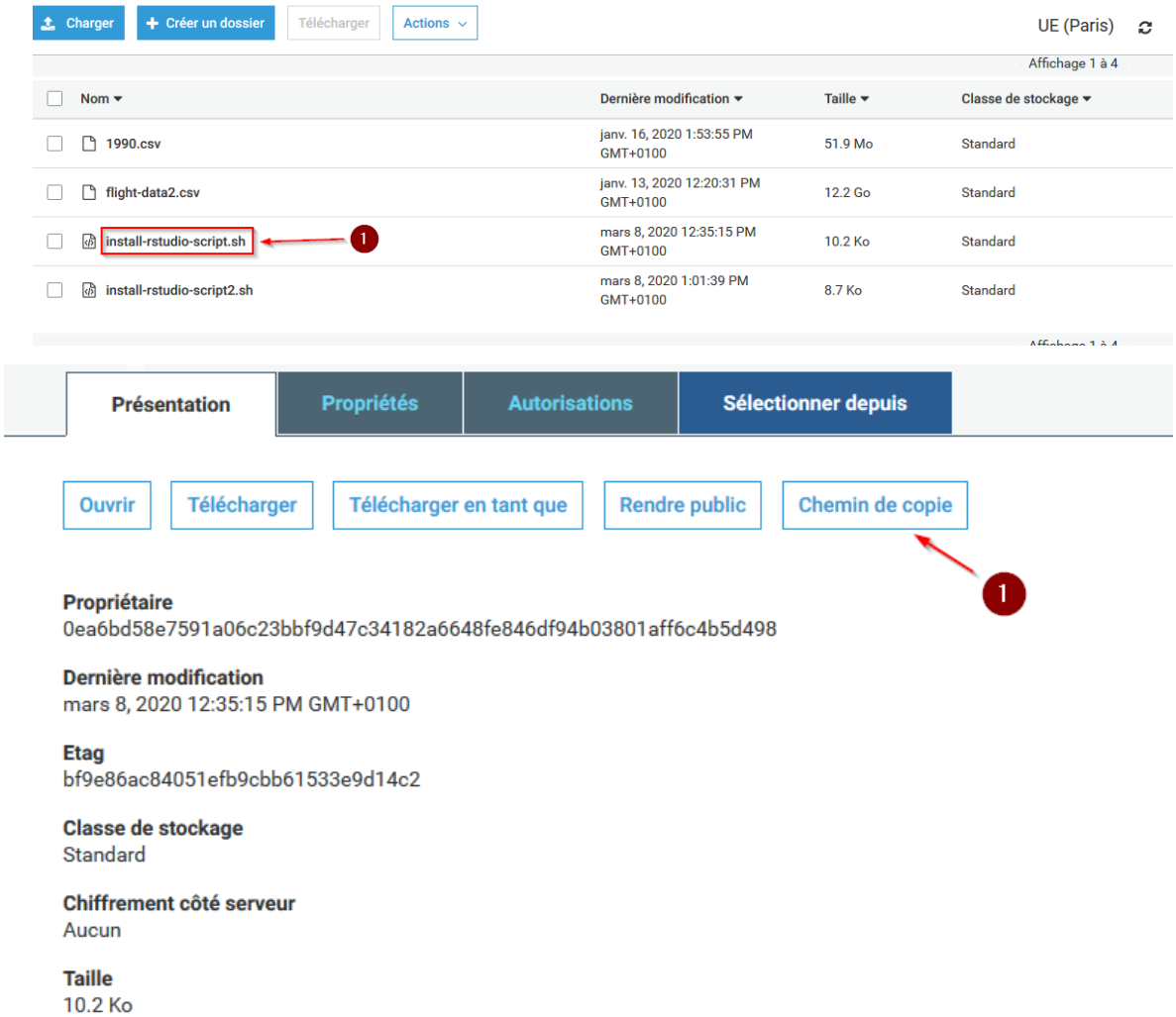
Installer Rstudio-server sur un cluster EMR via un script d'amorçage

Cette fiche contient les quelques étapes nécessaires pour installer Rstudio-server sur un cluster EMR en utilisant une action d'amorçage. Une action d'amorçage est une action réalisée au lancement du cluster EMR qui exécute un script sur tout les nœuds du cluster. Le script que vous allez utiliser va installer R, Rstudio-server et les différentes applications nécessaires ainsi de nombreuses bibliothèques R, ainsi vous n'aurez pas à les télécharger par la suite.

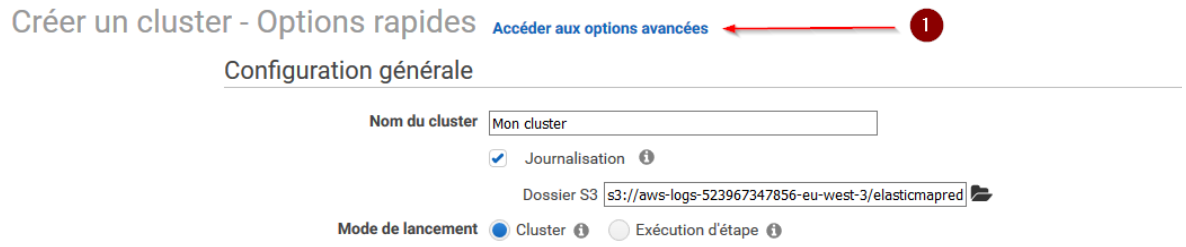
Copier le script d'amorçage sur S3

Vous trouverez le script d'amorçage sur Moodle, ou ici. Ce script n'est pas de nous, et provient de cet article. Il a néanmoins était modifié à la marge pour ne pas télécharger certaines applications inutiles. Déposez ensuite ce fichier dans un de vos compartiments S3. Puis allez cliquer sur "Chemin de copie" et copiez-le dans un bloc note.



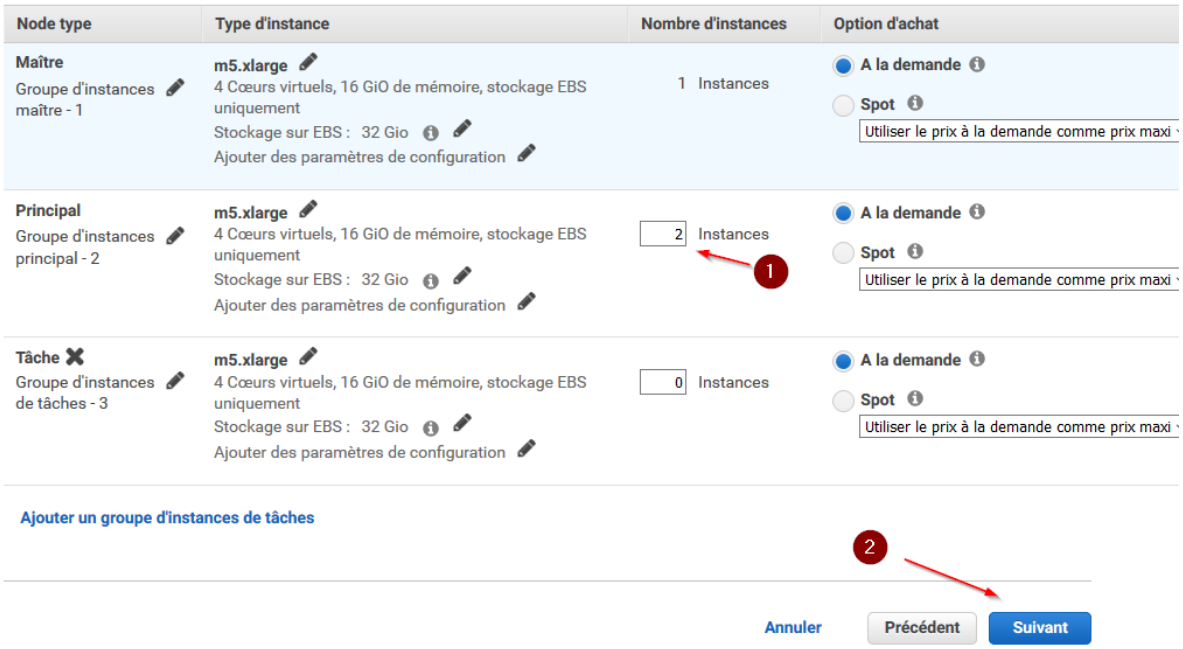
Créer un cluster EMR avec une action d'amorçage

Cliquez sur le bouton de création d'un cluster et accédez aux options avancées.

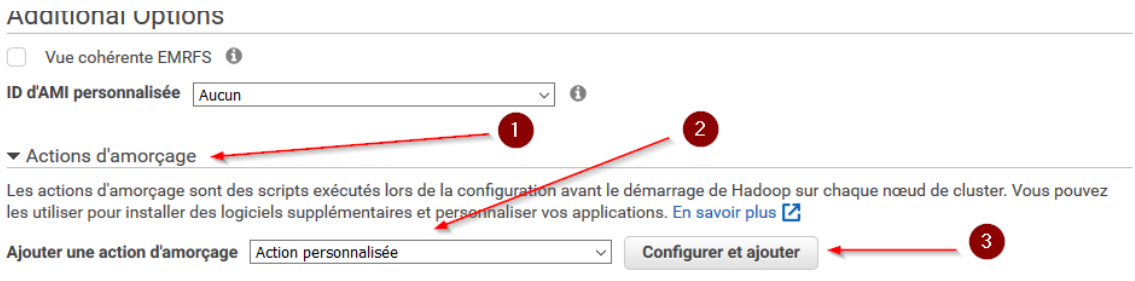


Choisissez la version emr-5.29.0 et les applications **Hadoop et Spark**. Vous pouvez en choisir d'autre comme Hive pour avoir accès à une base de données distribués. Puis passez à l'étape suivante

Sur l'étape suivante vous allez pouvoir définir les machines de votre cluster. Pas défaut il est constitué de 3 machines. Vous pouvez en augmenter le nombre (mais cela vous coutera plus cher). Passez à l'étape suivante.

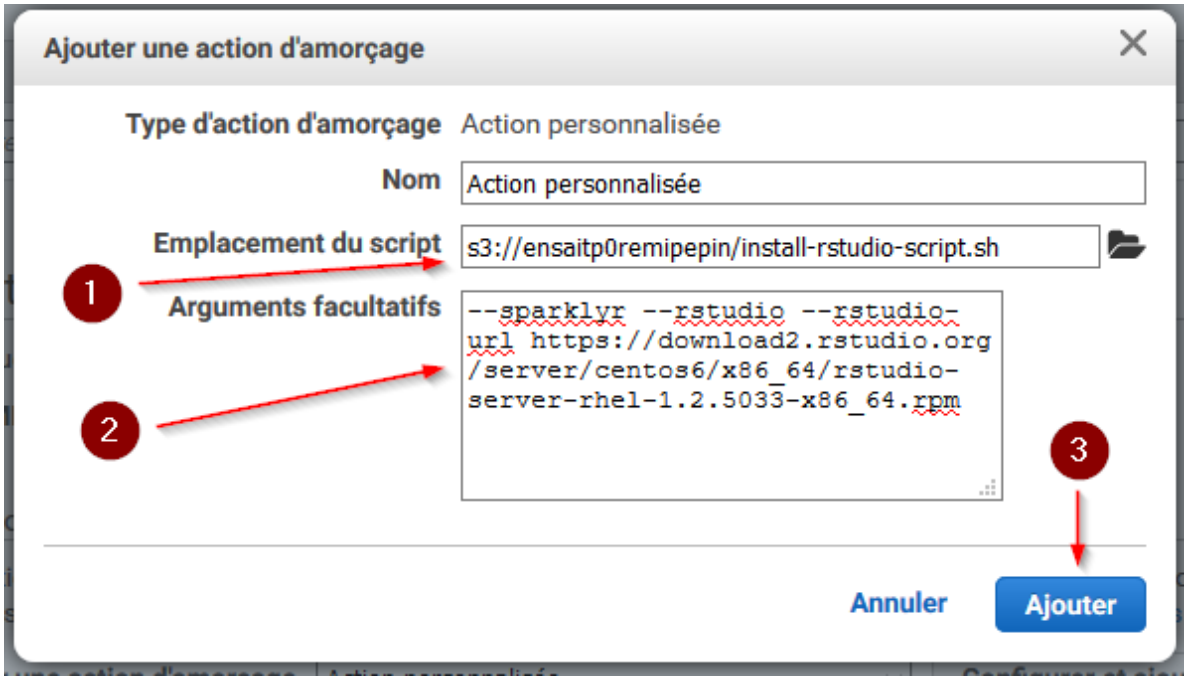


Sur la page suivante descendez en bas de la page, puis dépliez "Actions d'amorçage". Puis choisissiez "Action personnalisée" et "Configurer et ajouter"

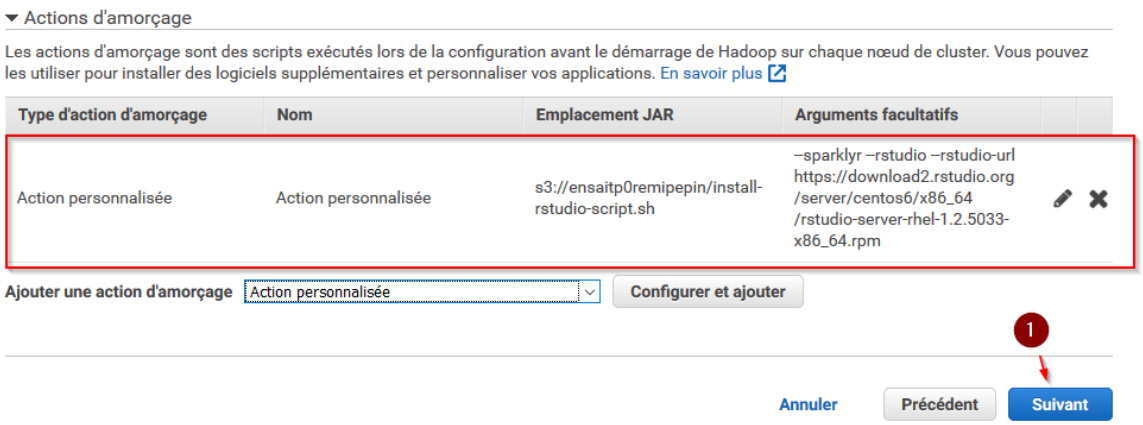


Maintenant vous allez configurer l'action d'amorçage. Pour l'emplacement du script copiez le chemin d'accès de votre script (le lien que vous devez avoir copier dans un bloc note). Puis rentrez les arguments suivants

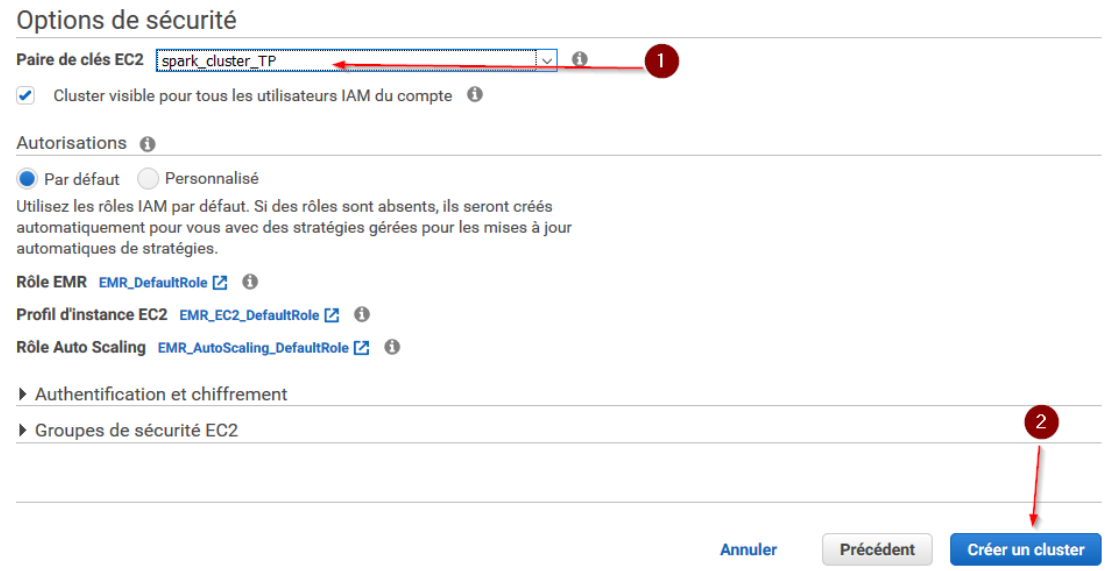
```
--sparklyr --rstudio --rstudio-url
https://download2.rstudio.org/server/centos6/x86_64/rstudio-server-rhel-
1.2.5033-x86_64.rpm
```



Une fois l'action ajoutée, une ligne va d'ajouter dans les actions d'amorçage. Passez à l'étape suivante.

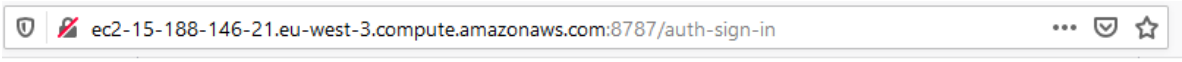


Sur le dernier écran choisissez votre clef ssh et créez le cluster. Le choix de la clef est obligatoire pour pouvoir se connecter au cluster par la suite.



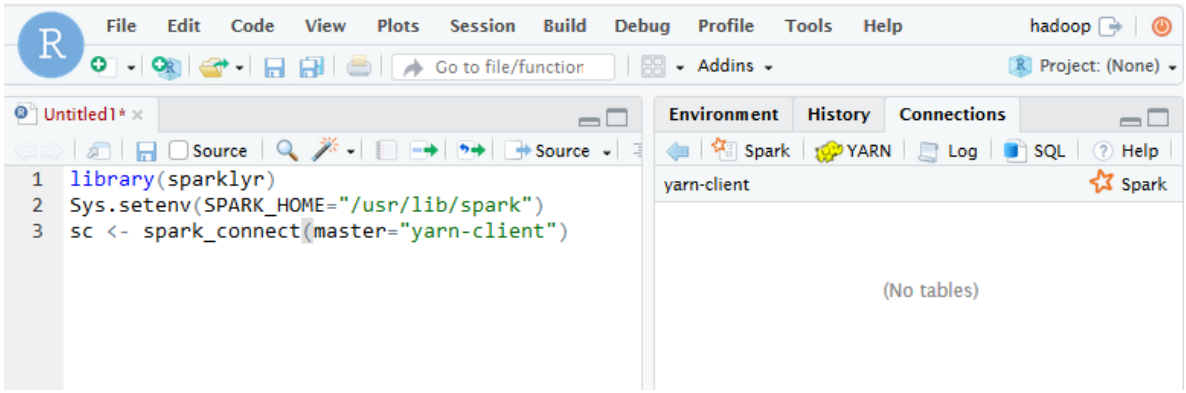
Le cluster va mettre quelques temps pour se créer et rester quelques minutes sur "Action d'amorçage". Cela est dû au fait que vous téléchargez R, Rstudio-server, et différentes bibliothèques R lourdes (comme sparklyR).

Une fois le cluster en état "En attente", connectez vous en SSH au cluster avec Putty (pensez à configurer la clef SSH, et le tunnel SSH) et vérifiez sur Foxy proxy est activé. Puis connecter vous à l'URL suivante http://dns.public.de.votre.cluster:8787, puis utiliser le user hadoop et le mot de passe hadoop pour vous connecter à Rstudio-server.



Ensuite connecter vous au cluster avec le code suivant :

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


File Edit Code View Plots Session Build Debug Profile Tools Help
hadoop
Go to file/function
Addins
Project: (None)
Untitled1*
Source
Source
1 library(sparklyr)
2 Sys.setenv(SPARK_HOME="/usr/lib/spark")
3 sc <- spark_connect(master="yarn-client")
Environment
History
Connections
Spark
YARN
Log
SQL
Help
yarn-client
Spark
(No tables)