

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.

Charger

+ Créer un dossier

Télécharger

Actions

UE (Paris)

Affichage 1 à 4

<input type="checkbox"/>	Nom	Dernière modification	Taille	Classe de stockage
<input type="checkbox"/>	1990.csv	janv. 16, 2020 1:53:55 PM GMT+0100	51.9 Mo	Standard
<input type="checkbox"/>	flight-data2.csv	janv. 13, 2020 12:20:31 PM GMT+0100	12.2 Go	Standard
<input type="checkbox"/>	install-rstudio-script.sh	mars 8, 2020 12:35:15 PM GMT+0100	10.2 Ko	Standard
<input type="checkbox"/>	install-rstudio-script2.sh	mars 8, 2020 1:01:39 PM GMT+0100	8.7 Ko	Standard

Affichage 1 à 4

Présentation

Propriétés

Autorisations

Sélectionner depuis

Ouvrir

Télécharger

Télécharger en tant que

Rendre public

Chemin de copie

Propriétaire

Oea6bd58e7591a06c23bbf9d47c34182a6648fe846df94b03801aff6c4b5d498

Dernière modification

mars 8, 2020 12:35:15 PM GMT+0100

Etag

bf9e86ac84051efb9cbb61533e9d14c2

Classe de stockage

Standard

Chiffrement côté serveur

Aucun

Taille

10.2 Ko

Créer un cluster EMR avec une action d'amorçage

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

Configuration générale

Nom du cluster

☒ Journalisation ⓘ

Dossier S3

Mode de lancement ☒ Cluster ⓘ ☐ Exécution d'étape ⓘ

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

Configuration des logiciels

Libérer

☒ Hadoop 2.8.5 ☐ Zeppelin 0.8.2 ☐ Livy 0.6.0

☐ JupyterHub 1.0.0 ☐ Tez 0.9.2 ☐ Flink 1.9.1

☐ Ganglia 3.7.2 ☐ HBase 1.4.10 ☐ Pig 0.17.0

☐ Hive 2.3.6 ☐ Presto 0.227 ☐ ZooKeeper 3.4.14

☐ MXNet 1.5.1 ☐ Sqoop 1.4.7 ☐ Mahout 0.13.0

☐ Hue 4.4.0 ☐ Phoenix 4.14.3 ☐ Oozie 5.1.0

☒ Spark 2.4.4 ☐ HCatalog 2.3.6 ☐ TensorFlow 1.14.0

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.

Node type	Type d'instance	Nombre d'instances	Option d'achat
Maître Groupe d'instances maître - 1	m5.xlarge 4 Cœurs virtuels, 16 GiO de mémoire, stockage EBS uniquement Stockage sur EBS : 32 Gio Ajouter des paramètres de configuration	1 Instances	<input checked="" type="radio"/> A la demande ⓘ <input type="radio"/> Spot ⓘ Utiliser le prix à la demande comme prix maxi
Principal Groupe d'instances principal - 2	m5.xlarge 4 Cœurs virtuels, 16 GiO de mémoire, stockage EBS uniquement Stockage sur EBS : 32 Gio Ajouter des paramètres de configuration	<input type="text" value="2"/> Instances	<input checked="" type="radio"/> A la demande ⓘ <input type="radio"/> Spot ⓘ Utiliser le prix à la demande comme prix maxi
Tâche ✕ Groupe d'instances de tâches - 3	m5.xlarge 4 Cœurs virtuels, 16 GiO de mémoire, stockage EBS uniquement Stockage sur EBS : 32 Gio Ajouter des paramètres de configuration	<input type="text" value="0"/> Instances	<input checked="" type="radio"/> A la demande ⓘ <input type="radio"/> Spot ⓘ Utiliser le prix à la demande comme prix maxi

[Ajouter un groupe d'instances de tâches](#)

Annuler

Précédent

Suivant

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

Additional options

☐ Vue cohérente EMRFS ⓘ

ID d'AMI personnalisée

▼ Actions d'amorçage

Les actions d'amorçage sont des scripts exécutés lors de la configuration avant le démarrage de Hadoop sur chaque nœud de cluster. Vous pouvez les utiliser pour installer des logiciels supplémentaires et personnaliser vos applications. [En savoir plus](#)

Ajouter une action d'amorçage

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

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

Ajouter une action d'amorçage

Type d'action d'amorçage: Action personnalisée

Nom: Action personnalisée

Emplacement du script: s3://ensaitp0remipepin/install-rstudio-script.sh

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

Annuler Ajouter

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

▼ Actions d'amorçage

Les actions d'amorçage sont des scripts exécutés lors de la configuration avant le démarrage de Hadoop sur chaque nœud de cluster. Vous pouvez les utiliser pour installer des logiciels supplémentaires et personnaliser vos applications. [En savoir plus](#)

Type d'action d'amorçage	Nom	Emplacement JAR	Arguments facultatifs
Action personnalisée	Action personnalisée	s3://ensaitp0remipepin/install-rstudio-script.sh	--sparklyr --rstudio --rstudio-url https://download2.rstudio.org/server/centos6/x86_64/rstudio-server-rhel-1.2.5033-x86_64.rpm

Ajouter une action d'amorçage: Action personnalisée

Configurer et ajouter

Annuler Précédent Suivant

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.

Options de sécurité

Paire de clés EC2 spark_cluster_TP 1

☒ Cluster visible pour tous les utilisateurs IAM du compte i

Autorisations i

☒ Par défaut ☐ Personnalisé

Utilisez les rôles IAM par défaut. Si des rôles sont absents, ils seront créés automatiquement pour vous avec des stratégies gérées pour les mises à jour automatiques de stratégies.

Rôle EMR [EMR_DefaultRole](#) i

Profil d'instance EC2 [EMR_EC2_DefaultRole](#) i

Rôle Auto Scaling [EMR_AutoScaling_DefaultRole](#) i

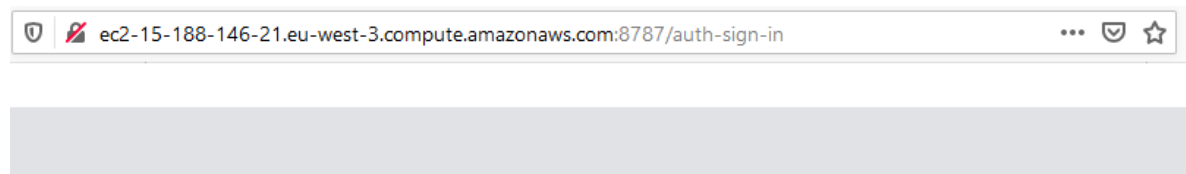
► Authentification et chiffrement

► Groupes de sécurité EC2 2

[Annuler](#) [Précédent](#) [Créer un cluster](#)

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.



Sign in to RStudio

Username:

Password:

☐ Stay signed in

[Sign In](#)

Ensuite connecter vous au cluster avec le code suivant :

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

The screenshot displays the RStudio IDE. The top menu bar contains 'File', 'Edit', 'Code', 'View', 'Plots', 'Session', 'Build', 'Debug', 'Profile', 'Tools', and 'Help'. The top right corner shows 'hadoop' and 'Project: (None)'. The left pane shows a file named 'Untitled1' with the following R code:

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

The right pane shows the 'Environment' tab with 'yarn-client' listed. Below it, it says '(No tables)'.