

Hadoop cluster creation in AWS

First: **DO NOT FORGET TO TURN YOUR CLUSTER OFF A THE END OF THIS TUTORIAL!**

- ☐ Go to AWS academy and the lab of the course (check your e-mail). Then go to **Modules** and **Learner Lab**
- ☐ Click on **Start Lab** then wait for the circle next to AWS to become green (should take 2-5 min)
- ☐ Click on **AWS Details** and download the lab's SSH key
- ☐ Click on **AWS** to go to AWS Mangement console.
- ☐ Once connected to the management console, search for "EMR" (Elastic Map Reduce). It a platform as a service made to manage Hadoop cluster in AWS. You just have to choose the configuration of your cluster (how many machines ? How many CPU/Ram ? Which release for Spark ?) and AWS will create your cluster. Doing this all by yourself is time consuming and not a pleasant task. That's why cloud providers provide service like EMR.
- ☐ You should land on a page like this

Amazon EMR

EMR on EC2

Clusters

Blocs-notes

Git repositories

Configurations de la sécurité

Bloquer l'accès public

Sous-réseaux VPC

Événements

EMR on EKS

Virtual clusters

Aide

Nouveautés

Bienvenue dans Amazon Elastic MapReduce

Amazon Elastic MapReduce (Amazon EMR) est un service Web qui permet aux commerces, aux chercheurs, aux analystes de données et aux développeurs de traiter de grandes quantités de données de manière simple et économique.

Apparemment, vous n'avez aucun cluster. Créez-en un maintenant :

Créer un cluster

Fonctionnement d'Elastic MapReduce

Charger



Téléchargez les données et l'application de traitement sur S3.

Créer



Configurez et créez votre cluster en spécifiant les entrées et les sorties de données, la taille de cluster, etc.

Contrôle



Surveillez l'état et la progression de votre cluster. Récupérez la sortie d'une tâche.

Informations supplémentaires

En savoir plus sur Elastic MapReduce :

- [Présentation d'EMR](#)
- [FAQ](#)
- [Tarification](#)

En savoir plus sur l'utilisation d'Elastic MapReduce :

- [Forum](#)
- [Documentation](#)
- [Manuel du développeur](#)
- [Référence API](#)
- [EMR sur Github](#)
- [Portail d'aide](#)

Next time it should be this one.

Amazon EMR

EMR on EC2

Clusters

Blocs-notes

Git repositories

Configurations de la sécurité

Bloquer l'accès public

Sous-réseaux VPC

Événements

EMR on EKS

Virtual clusters

Aide

Nouveautés

Créer un cluster

Afficher les détails

Cloner

Résilier

Filtre : Tous les clusters

4 clusters (tous chargés)

	Nom	ID	Statut	Heure de création (UTC+1)	Temps écoulé	Heures normales
<input type="checkbox"/>	Mon cluster	j-1H72GTU5MWKU2	Résilié Demande utilisateur	23-03-2021 15:23 (UTC+1)	53 minutes	24
<input type="checkbox"/>	Mon cluster	j-38AR1CNY33PPB	Résilié Demande utilisateur	23-03-2021 15:15 (UTC+1)	11 minutes	24
<input type="checkbox"/>	NotebookCluster	j-39CG54L47FWG	Résilié Demande utilisateur	23-03-2021 14:52 (UTC+1)	13 minutes	8
<input type="checkbox"/>	Mon cluster	j-2P3UZ7CHZ3BHC	Résilié Demande utilisateur	23-03-2021 14:44 (UTC+1)	9 minutes	24

In every cases click on **Cluster** then **Create Cluster**

Amazon EMR

EMR on EC2

Clusters

Blocs-notes

Git repositories

Configurations de la sécurité

Bloquer l'accès public

Sous-réseaux VPC

Événements

EMR on EKS

Virtual clusters

Aide

Blocs-notes

Utilisez des blocs-notes Jupyter gérés par EMR pour analyser les données de façon interactive avec du code en direct, des textes narratifs, des visualisations, et bien plus. Les blocs-notes sont exécutés gratuitement et sont enregistrés dans Amazon S3 indépendamment des clusters. La facturation standard pour les clusters et Amazon EMR s'applique.

Créer un bloc-notes Afficher les détails Ouvrir dans JupyterLab Ouvrir dans Jupyter Démarrage de Arrêter Supprimer

Filtre : Tous les blocs-notes Filtrer les blocs-notes... 1 bloc-notes (tout chargé)

Nom	Statut	Clus
test	Arrêté	1H

- ☐ You notebook configuration should be
- ☐ **Cluster Name** : a simple name like "my hadoop cluster"
 - ☐ Launch type **cluster** for a long living cluster
 - ☐ **Release** : choose the latest release (6.9.0) and **Core Hadoop** for the installed application
 - ☐ For the instance type keep the default m5.xlarge. If need you can select c4.xlarge, m4.xlarge, c5.xlarge or r4.rlarge.
 - ☐ Select the **vockey** for the EC2 key pair.
 - ☐ Then **Create cluster**

General Configuration

Cluster name

☒ Logging ⓘ

S3 folder

Launch mode ☒ Cluster ⓘ ☐ Step execution ⓘ

Can be anything

Software configuration

Release ⓘ

Applications

- ☒ Core Hadoop: Hadoop 3.3.3 with Hive 3.1.3, Hue 4.10.0, Pig 0.17.0 and Tez 0.10.2
- ☐ HBase: HBase 2.4.13 with Hadoop 3.3.3, Hive 3.1.3, Hue 4.10.0, Phoenix 5.1.2, and ZooKeeper 3.5.10
- ☐ Presto: Presto 0.276 with Hadoop 3.3.3 HDFS and Hive 3.1.3 Metastore
- ☐ Spark: Spark 3.3.0 on Hadoop 3.3.3 HDFS and Zeppelin 0.10.1
- ☐ Trino: Trino 398 with Hadoop 3.3.3 HDFS and Hive 3.1.3 Metastore

☐ Use AWS Glue Data Catalog for table metadata ⓘ

Choose emr-6.9.0 and core hadoop for a basic hadoop cluster.

Hardware configuration

Instance type

Number of instances (1 master and 2 core nodes)

Cluster scaling ☐ scale cluster nodes based on workload

Auto-termination ☒ Enable auto-termination [Learn more](#) ⓘ

Terminate cluster when it is idle after hours

The default m5.xlarge works fine, but if there are issues with instance disponibility you can use m4.xlarge, c4.xlarge, c5.xlarge or r4.rlarge. Keep 3 for the cluster size Change the max idle time

Security and access

EC2 key pair ⓘ [Learn how to create an EC2 key pair.](#)

Permissions ☒ Default ☐ Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role [EMR_DefaultRole](#) ⓘ ☐ Use EMR_DefaultRole_V2 ⓘ

EC2 instance profile [EMR_EC2_DefaultRole](#) ⓘ

Choose the vockey key pair

[Cancel](#) [Create cluster](#)

- ☐ ⌚ The cluster creation takes time (between 5 and 10min), please wait and read this tutorial.

Here is a table with the hourly price of some instances just to give you an idea of the cost of an EMR cluster (hourly instance price*cluster size)

Instance	Hourly price per instance
m4.xlarge	0.24 \$
m5.xlarge	0.23\$
c4.xlarge	0.25\$
c5.xlarge	0.22\$
r4.xlarge	0.30\$
c5.24xlarge	5,3\$

- ☐ Once create your page should looks like that. Congrats you cluster is up and alive ! But for security reasons, your cluster is currently unreachable. It's not the cluster fault, but its security group. To access it we have to allow SSH connection with the security groupe.
- ☐ Right click on `Security Group for Master` and open the link in another tab to make you cluster accessible with SSH connection.

The screenshot shows the AWS EMR console interface. At the top, there are buttons for 'Clone', 'Terminate', and 'AWS CLI export'. Below them, the cluster name 'my hadoop cluster' is shown with a 'Waiting' status and a message 'Cluster ready to run steps.' A navigation bar includes tabs for 'Summary', 'Application user interfaces', 'Monitoring', 'Hardware', 'Configurations', 'Events', 'Steps', and 'Bootstrap actions'. The 'Summary' tab is active, displaying details like ID, creation date, and configuration. The 'Security and access' section at the bottom contains a red box around the 'Security groups for Master' link, with a red arrow pointing to it from the right.

- ☐ On the next page select the security group name `ElasticMapReduce-master`, in the bottom of the screen select `inbound rules` and finally `Edit inbound rules`
- ☐ Scroll all the way down and select `Add rule`. For the new rule select `SSH` then `My IP`. Then `save rules`
- ☐ Go back to your cluster page, click on `Connect to the Maste Node ussing SSH` and follow the instruction. Basically you will just do a basic SSH connection, with the target host being your master node using its Master public DNS and the user is `hadoop`, and using the previously downloaded key (step 3).

