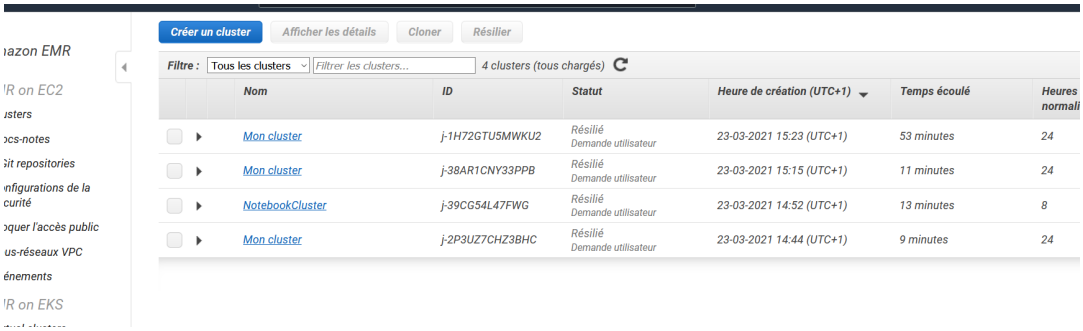
Hadoop cluster creation in AWS

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

Next time it should be this one.



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

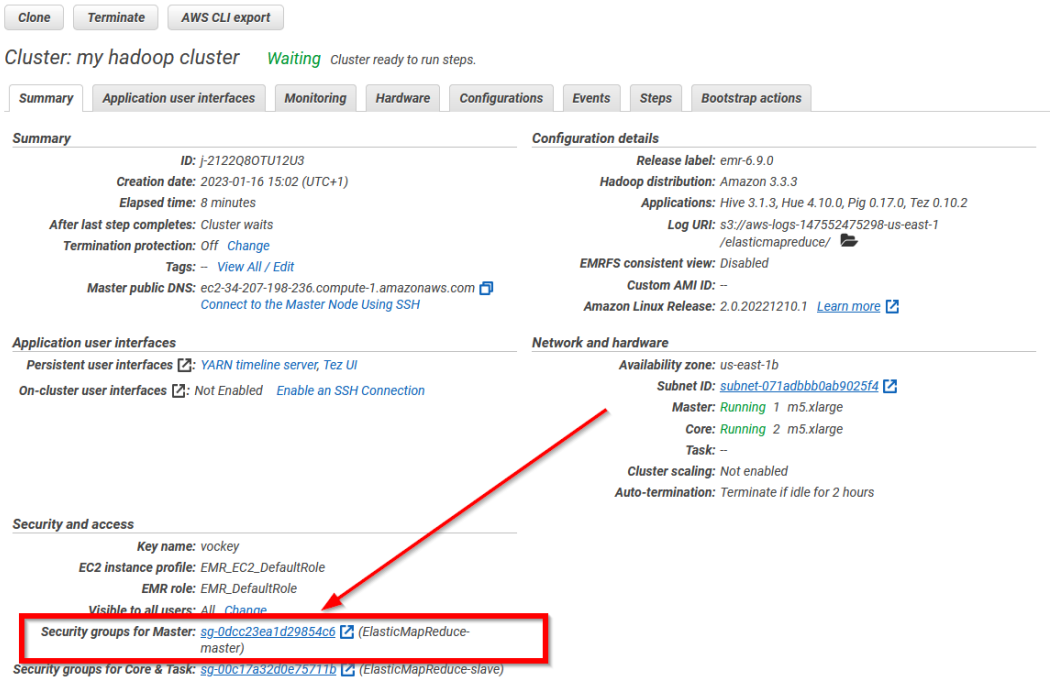
- [ ] You notebook configuration should be
  - [ ] `Cluster Name` : a simple name like "my hadoop cluster"
  - [ ] Lauch type `cluster` for a long living cluster
  - [ ] `Release` : choose the latest release (6.9.0) and `Core Hadoop` for the installed application
  - [ ] For the instance type keep the 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`

- [ ] ⏳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.



- [ ] 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).