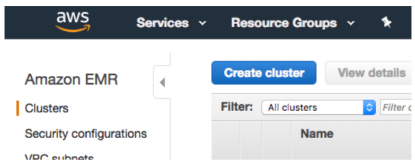
Launch EMR Cluster and Notebook

Follow the instructions below to launch your EMR cluster and notebook.

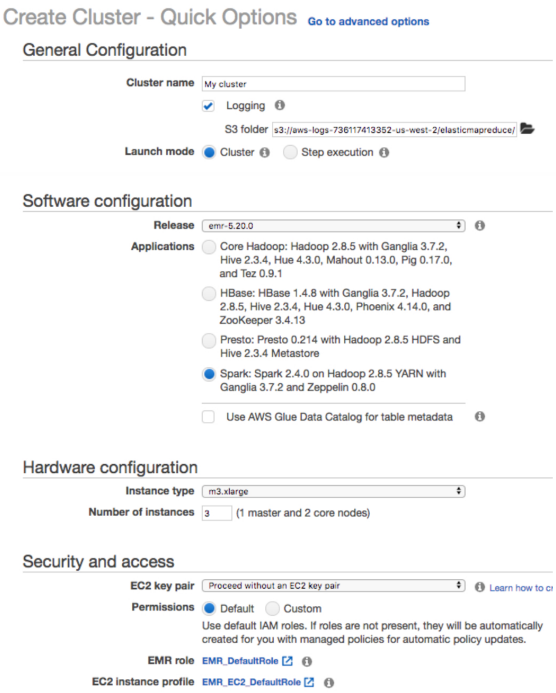
- Go to the Amazon EMR Console
- Select "Clusters" in the menu on the left, and click the "Create cluster" button.



Step 1: Configure your cluster with the following settings:

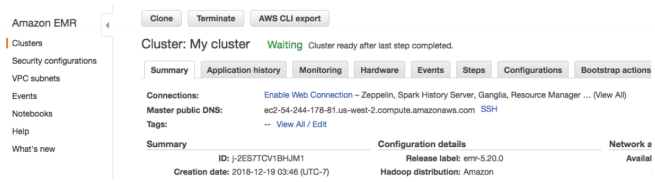
- Release: `emr-5.20.0` or later
- Applications: `Spark`: Spark 2.4.0 on Hadoop 2.8.5 YARN with Ganglia 3.7.2 and Zeppelin 0.8.0
- Instance type: `m3.xlarge`
- Number of instance: `3`
- EC2 key pair: `Proceed without an EC2 key pair` or feel free to use one if you'd like

You can keep the remaining default setting and click "Create cluster" on the bottom right.



Step 2: Wait for Cluster "Waiting" Status

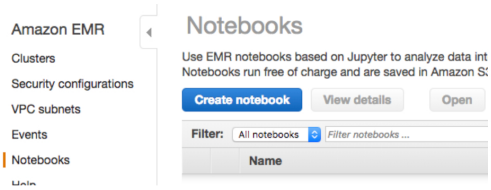
Once you create the cluster, you'll see a status next to your cluster name that says *Starting*. Wait a short time for this status to change to *Waiting* before moving on to the next step.



Step 3: Create Notebook

Now that you launched your cluster successfully, let's create a notebook to run Spark on that cluster.

Select "Notebooks" in the menu on the left, and click the "Create notebook" button.

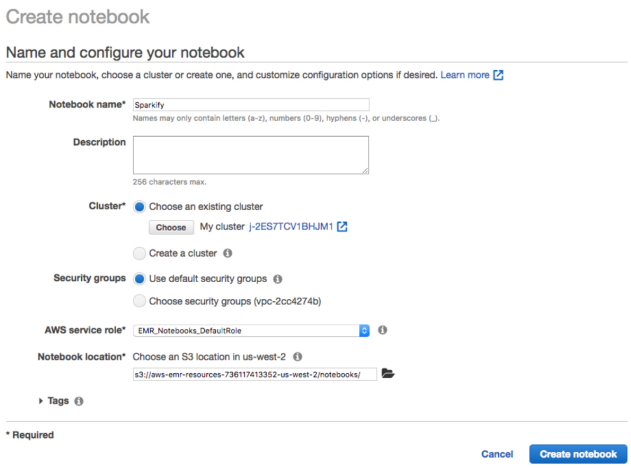


Step 4: Configure your notebook

Step 4: Configure your Notebook

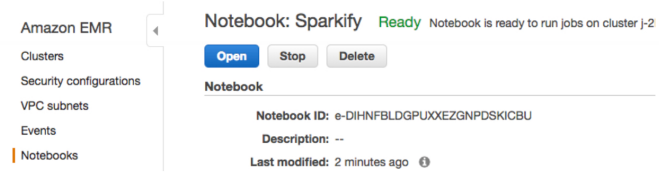
- Enter a name for your notebook
- Select "Choose an existing cluster" and choose the cluster you just created
- Use the default setting for "AWS service role" - this should be "EMR_Notebooks_DefaultRole" or "Create default role" if you haven't done this before.

You can keep the remaining default settings and click "Create notebook" on the bottom right.



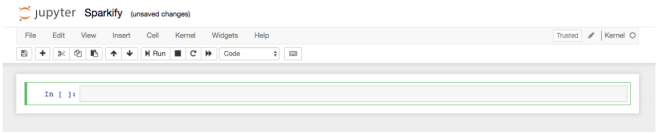
Step 5: Wait for Notebook "Ready" Status, Then Open

Once you create an EMR notebook, you'll need to wait a short time before the notebook status changes from *Starting* or *Pending* to *Ready*. Once your notebook status is *Ready*, click the "Open" button to open the notebook.



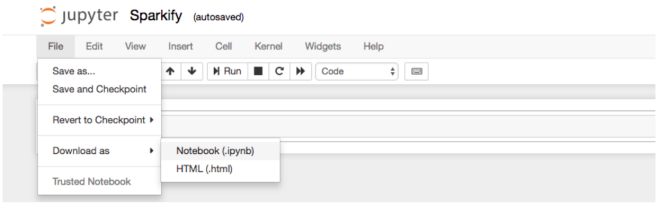
Start Coding!

Now you can run Spark code for your project in this notebook, which EMR will run on your cluster. In the next page, you'll find starter code to create a spark session and read in the full 12GB dataset for the DSND Capstone project.



Download Notebook

When you are finished with your notebook, click `File` > `Download as` > `Notebook` to download it to your computer. On your local computer, create a git repository including this notebook and a README file. Submit the URL to your github repository to submit this project. See more details in the Sparkify Project Overview page.



For more information on EMR notebooks, click here.

Pricing - Be Careful!

From this point on, AWS will charge you for running your EMR cluster. See details on this and how to manage your resources to avoid unexpected costs in the "Managing Resources" section at the end of this lesson.

NEXT