1: Introduction

Link: <https://www.youtube.com/watch?v=kv08_VWQ2Ho>

Notes:

2: Lesson overview

Link: <https://www.youtube.com/watch?v=fLzYJsJozvM>

Notes:

3: Why data lakes: Evolution of the data warehouse

Link: <https://www.youtube.com/watch?v=3KSt8-nFUo0>

Notes:

4: Why data lakes: unstructured and big data

Link: <https://www.youtube.com/watch?v=niF5i3qsagM>

Notes:

5: Why data lakes: new roles and advanced analytics

Link: <https://www.youtube.com/watch?v=3B7iv1GRaAw>

Notes:

6: Big data effects: low costs, ETL offloading

Link: <https://www.youtube.com/watch?v=-snvU7S3wZo>

Notes:

7: Big data effects: Schema-on-read

Link: <https://www.youtube.com/watch?v=cBvKpHfTLTA>

Notes:

8: Big data effects: (Un-/Semi-) Structured support

Link: <https://www.youtube.com/watch?v=5R8y9Xa8izs>

Notes:

9: Demo: Schema on read Pt 1

Link: <https://www.youtube.com/watch?v=_clMPKkFytg>

Notes:

10: Demo: Schema on read Pt 2

Link: <https://www.youtube.com/watch?v=_GUDMCEaX0Y>

Notes:

11: Demo: Schema on read Pt 3

Link: <https://www.youtube.com/watch?v=EN0AJPNjyeE>

Notes:

12: Demo: Schema on read Pt 4

Link: <https://www.youtube.com/watch?v=m1kD8EyFPH0>

Notes:

13: Exercise 1: Schema on read

Link:

Notes:

14: Demo: Advanced Analytics NLP Pt 1

Link: <https://www.youtube.com/watch?v=_ZMWq5cZjzI>

Notes:

15: Demo: Advanced Analytics NLP Pt 2

Link: <https://www.youtube.com/watch?v=gx3JYVDitiE>

Notes:

16: Demo: Advanced Analytics NLP Pt 3

Link: <https://www.youtube.com/watch?v=Ymgj8KetT14>

Notes:

17: Exercise: Advanced Analytics NLP

Link:

Notes:

18: Data Lake Implementation Introduction

Link: <https://www.youtube.com/watch?v=wi0OXQpr81U>

Notes:

19: Data lake concepts

Link: <https://www.youtube.com/watch?v=DozXTpjZjas>

Notes:

20: Data lake vs data warehouse

Link: <https://www.youtube.com/watch?v=2YkwP4jw8ig>

Notes:

21: Data lake options on AWS

Link: <https://www.youtube.com/watch?v=AMdFTlYWmp8>

Notes:

22: AWS Options: EMR (HDFS + Spark)

Link: <https://www.youtube.com/watch?v=m88kdMKB4qs>

Notes:

23: AWS options: EMR: S3 + Spark

Link: <https://www.youtube.com/watch?v=ur1clGbA0Rw>

Notes:

24: AWS options: Athena

Link: <https://www.youtube.com/watch?v=XNEczLKSISw>

Notes:

25: Demo: Data lake on S3 Pt 1

Link: <https://www.youtube.com/watch?v=phZ2irpPtMM>

Notes:

26: Demo: Data lake on S3 Pt 2

Link: <https://www.youtube.com/watch?v=kqy6w6kat58>

Notes:

27: Exercise: Data lake on S3

Link:

Notes:

28: Demo: Data lake on EMR Pt 1

Link: <https://www.youtube.com/watch?v=539RnL05fGQ>

Notes:

29: Demo: Data lake on EMR Pt 2

Link: <https://www.youtube.com/watch?v=lr-6oUkzs2w>

Notes:

30: Demo: Data lake on Athena Pt 1

Link: <https://www.youtube.com/watch?v=Cy-kJEc5oKE>

Notes:

31: Demo: Data lake on Athena Pt 2

Link: <https://www.youtube.com/watch?v=LWOTnS2wS9U>

Notes:

32: Data lake issues

Link: <https://www.youtube.com/watch?v=FCXdgt-IkTE>

Notes:

33: AWS: launch EMR cluster and Notebook

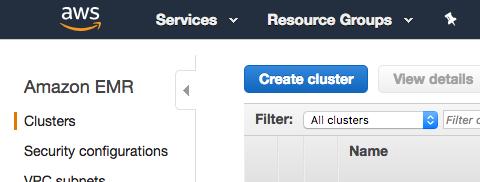
Link:

Notes:

# Launch EMR Cluster and Notebook

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

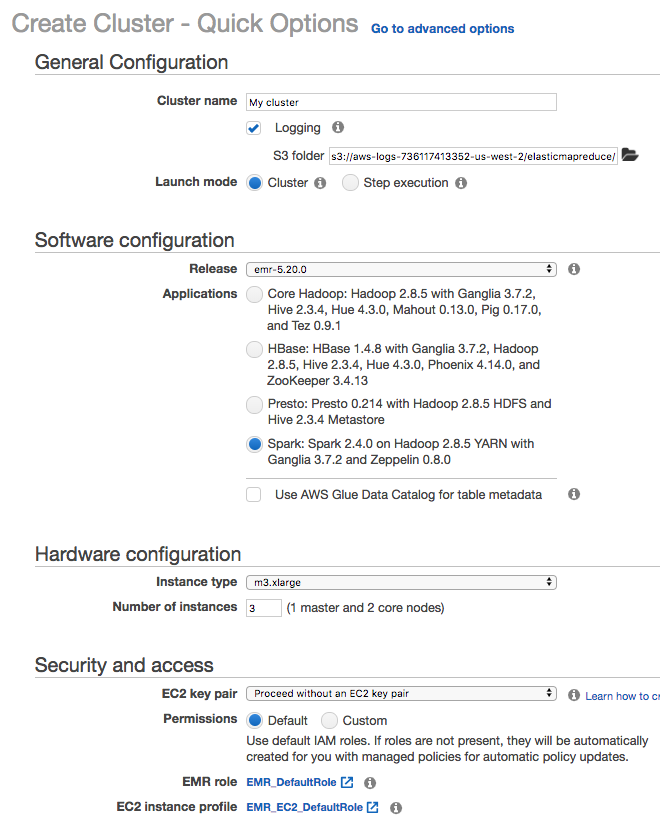
* Open a regular AWS account (if you don't already have one) following the instructions via the [**Amazon Web Service Help Center**](https://aws.amazon.com/premiumsupport/knowledge-center/create-and-activate-aws-account/)
* Go to the [**Amazon EMR Console**](https://console.aws.amazon.com/elasticmapreduce/)
* Select "Clusters" in the menu on the left, and click the "Create cluster" button.



# Step 1: Configure your cluster with the following settings:

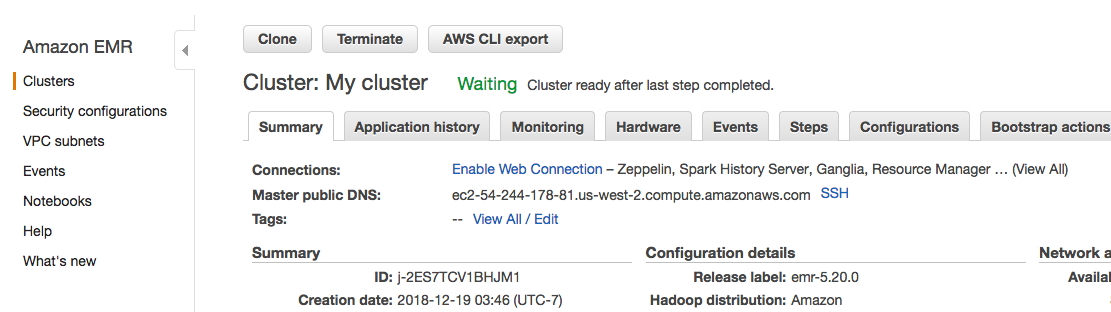
* Release: emr-5.20.0 or later
* Applications: Spark: Spark 2.4.0 on Hadoop 2.8.5 YARN with Ganglia 3.7.2 and Zeppelin 0.8.0
* Instance type: m3.xlarge
* Number of instance: 3
* EC2 key pair: Proceed without an EC2 key pair or feel free to use one if you'd like

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



# Step 2: Wait for Cluster "Waiting" Status

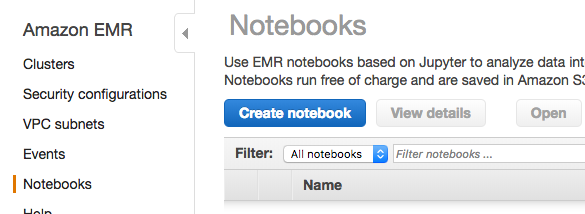
Once you create the cluster, you'll see a status next to your cluster name that says Starting. Wait a short time for this status to change to Waiting before moving on to the next step.



# Step 3: Create Notebook

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

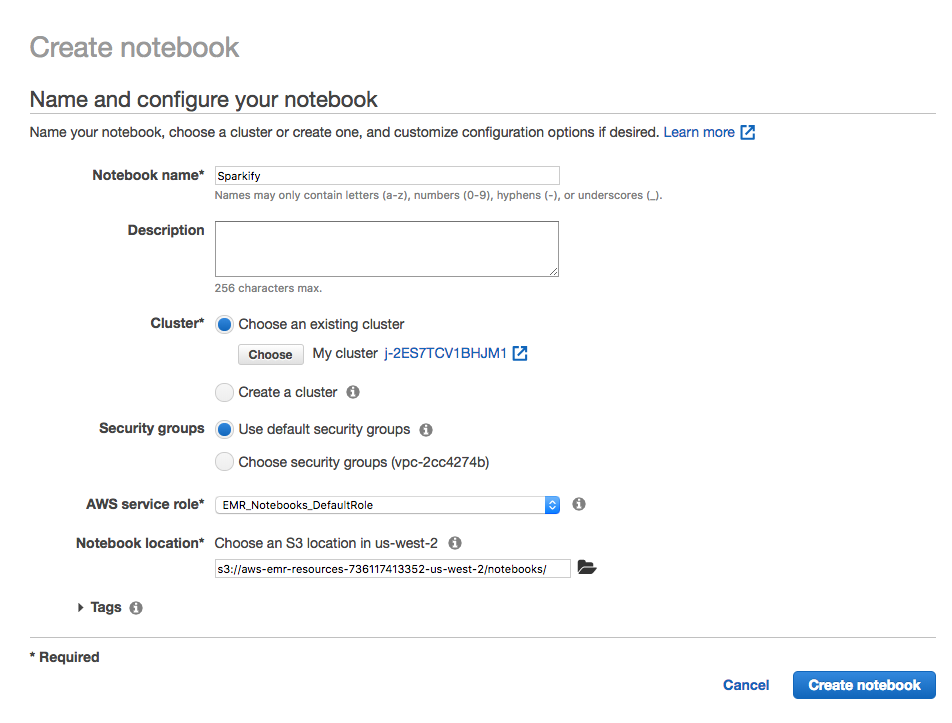
Select "Notebooks" in the menu on the left, and click the "Create notebook" button.



# Step 4: Configure your notebook

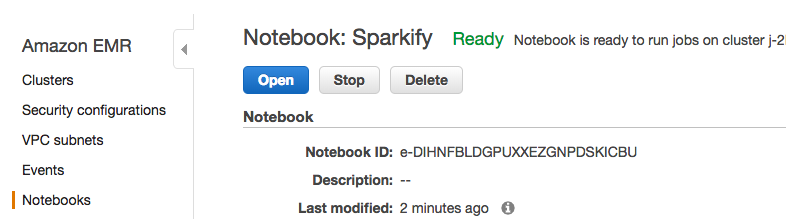
* Enter a name for your notebook
* Select "Choose an existing cluster" and choose the cluster you just created
* Use the default setting for "AWS service role" - this should be "EMR\_Notebooks\_DefaultRole" or "Create default role" if you haven't done this before.

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



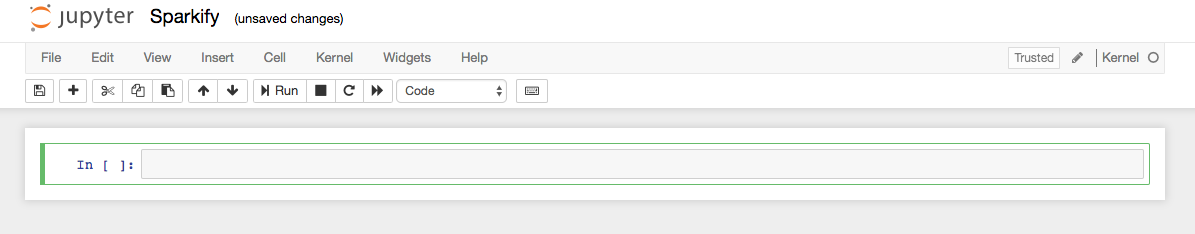
# Step 5: Wait for Notebook "Ready" Status, Then Open

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.



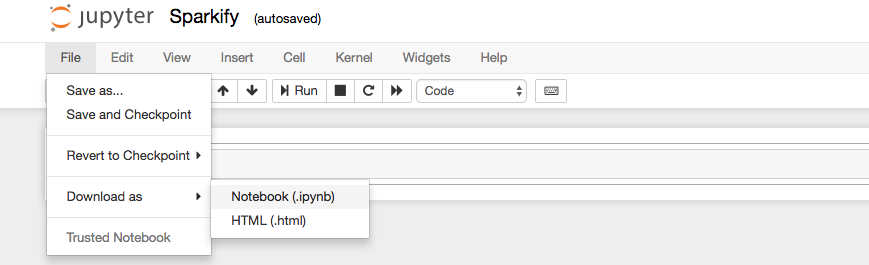
# Start Coding!

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](https://classroom.udacity.com/nanodegrees/nd025/parts/84260e1f-2926-4127-895f-cc4432b05059/modules/80c955ce-72f2-403a-9bf5-cc58636dab9d/lessons/d6285247-6bc0-4783-b118-6f41981b9469/concepts/47d96c80-82da-4640-9fff-54415f2a21df" \t "_blank)**.



For more information on EMR notebooks, click [**here**](https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-managed-notebooks.html).

# 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.

34: AWS: Avoid paying unexpected costs

Link:

Notes:

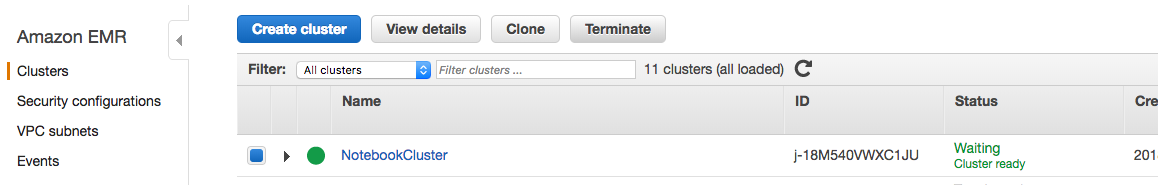
# Pricing - Be Careful!

From this point on, AWS will charge you for running your EMR cluster. You can find the details on the **[EMR Pricing here](https://aws.amazon.com/emr/pricing/" \t "_blank)**. If you run your cluster for a week with the settings specified earlier (3 instances of m3.xlarge), you should expect costs to be around $30, which should be covered in the amount of free promotional credits you have received from AWS as a Udacity student.

Most importantly, remember to terminate your cluster when you are done. Otherwise, your cluster might run for a day, week, month, or longer without you remembering, and you’ll wind up with a large bill!

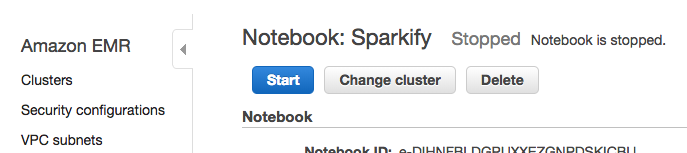
# Terminate Your Cluster

You can terminate your cluster while still keeping the Jupyter notebook you created. In EMR, your EMR cluster and EMR notebook are decoupled (so you can reattach your notebook to a different cluster at any time)! To terminate your cluster, click "Clusters" in the menu on the left, check the box next to your cluster to select, and click the "Terminate" button.

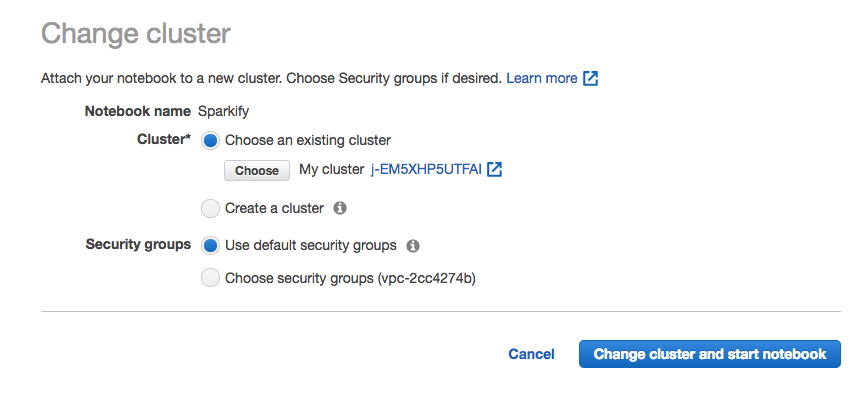


# Change Cluster for Notebook

If you still have a notebook on EMR and terminated the cluster it was connected to, you can still run that notebook at any time by creating another cluster (following the instructions in the previous section). Once you have the new cluster launched and in "waiting" status, click "Notebooks" in the menu on the left and click on the name of your notebook. Then click the "Change cluster" button.

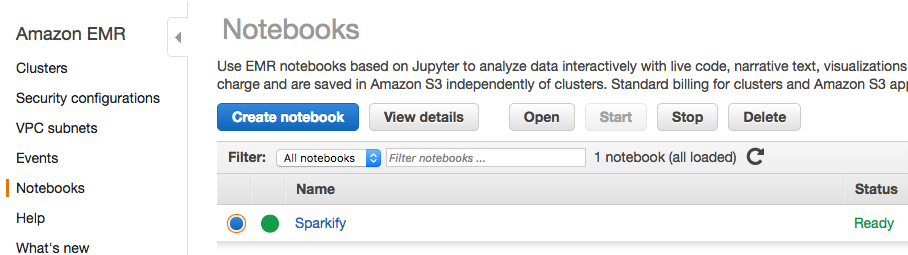


Select your new cluster. Once your notebook reaches "Ready" status, you can now run this notebook on your new cluster.



# Delete Your Notebook

When you've finished with your project and downloaded your notebook, you can delete your notebook from EMR by selecting "Notebooks" in the menu on the left, selecting your notebook, and then clicking "Delete." Make sure to terminate the clusters you launched for this as well.



# Delete S3 buckets

AWS charges primarily for running instances, so most of the charges will cease once you stop the cluster. However, there are smaller storage charges that continue to accrue if you don't delete your buckets. To delete your buckets, go to the **[Amazon S3 console](https://console.aws.amazon.com/s3/" \t "_blank)**. Choose the bucket you want to delete from the list, so that the whole bucket row is selected. Choose delete bucket, type the name of the bucket, and click "Confirm."

For more information about deleting folders and buckets, go to **[How Do I Delete an S3 Bucket](https://docs.aws.amazon.com/AmazonS3/latest/user-guide/delete-bucket.html" \t "_blank)** in the Amazon Simple Storage Service Getting Started Guide.

You can view your billing information at any time by clicking on your account name on the upper right corner of the console and go to **My Billing Dashboard**.

