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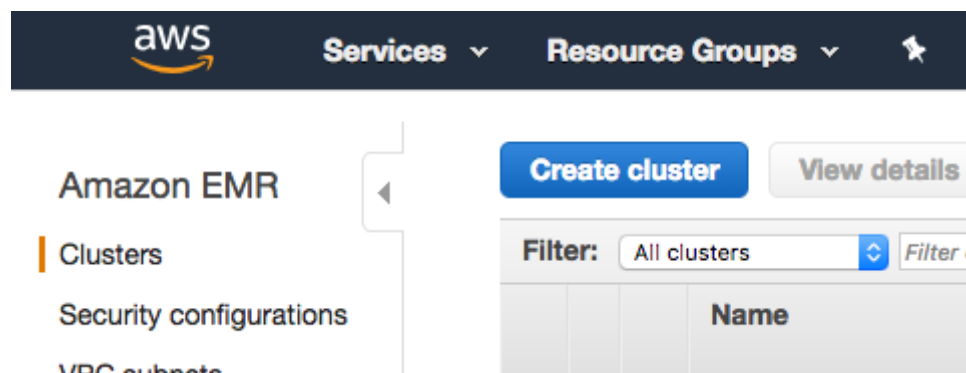
Follow the instructions below to launch your EMR cluster and notebook.

- 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.

Create Cluster - Quick Options   Go to advanced options

General Configuration

Cluster name   My cluster

☑ Logging ℹ

S3 folder   s3://aws-logs-736117413352-us-west-2/elasticmapreduce/ 📁

Launch mode   🔵 Cluster ℹ   ⚪ Step execution ℹ

Software configuration

Release   emr-5.20.0   ⇕ ℹ

Applications
⚪ Core Hadoop: Hadoop 2.8.5 with Ganglia 3.7.2, Hive 2.3.4, Hue 4.3.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.1

⚪ HBase: HBase 1.4.8 with Ganglia 3.7.2, Hadoop 2.8.5, Hive 2.3.4, Hue 4.3.0, Phoenix 4.14.0, and ZooKeeper 3.4.13

⚪ Presto: Presto 0.214 with Hadoop 2.8.5 HDFS and Hive 2.3.4 Metastore

🔵 Spark: Spark 2.4.0 on Hadoop 2.8.5 YARN with Ganglia 3.7.2 and Zeppelin 0.8.0

☐ Use AWS Glue Data Catalog for table metadata ℹ

Hardware configuration

Instance type   m3.xlarge   ⇕

Number of instances   3   (1 master and 2 core nodes)

Security and access

EC2 key pair   Proceed without an EC2 key pair   ⇕   ℹ Learn how to cr

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 🔗 ℹ

EC2 instance profile   EMR_EC2_DefaultRole 🔗 ℹ

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.

Create notebook

Name and configure your notebook

Name your notebook, choose a cluster or create one, and customize configuration options if desired. Learn more ⬀

Notebook name* Sparkify

Names may only contain letters (a-z), numbers (0-9), hyphens (-), or underscores (_).

Description

256 characters max.

Cluster* ● Choose an existing cluster

Choose   My cluster j-2ES7TCV1BHJM1 ⬀

○ Create a cluster ⓘ

Security groups ● Use default security groups ⓘ

○ Choose security groups (vpc-2cc4274b)

AWS service role* EMR_Notebooks_DefaultRole ⓘ

Notebook location* Choose an S3 location in us-west-2 ⓘ

s3://aws-emr-resources-736117413352-us-west-2/notebooks/ 📁

▸ **Tags** ⓘ

* Required

Cancel   **Create notebook**

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.

Amazon EMR ◂

Clusters

Security configurations

VPC subnets

Events

| Notebooks

Notebook: Sparkify Ready Notebook is ready to run jobs on cluster j-2

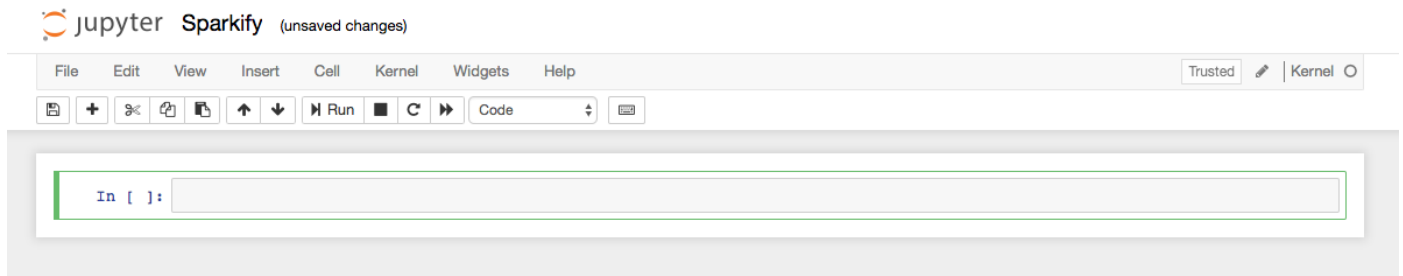**Open**   **Stop**   **Delete**

Notebook

Notebook ID: e-DIHNFBLDGPUXXEZGNPDSKICBU

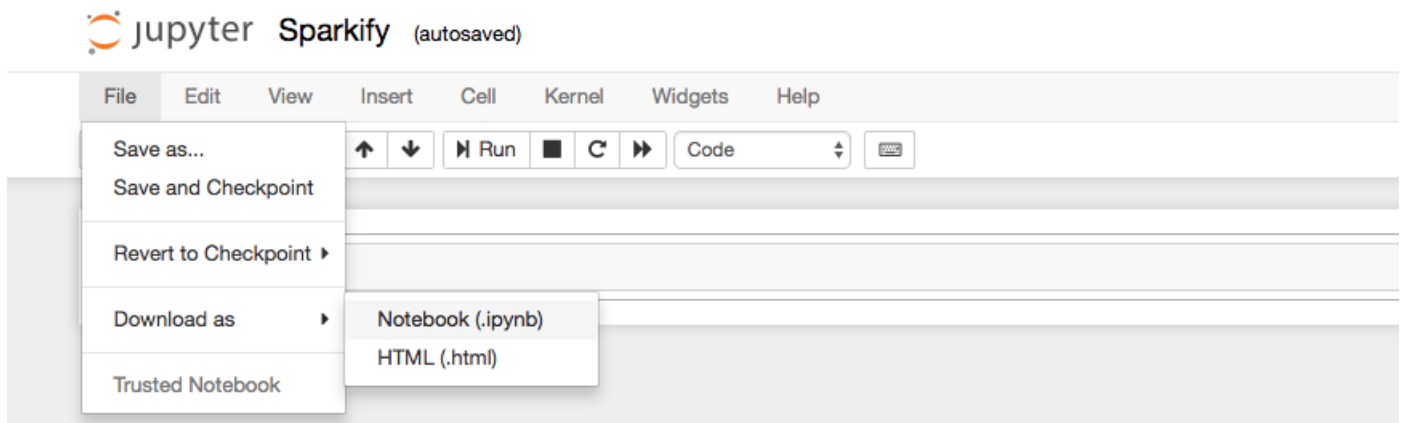**Description:** --

Last modified: 2 minutes ago ⓘ

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.