Spark on Amazon EMR (for CS 205)

**Very Important: Please terminate the cluster as soon as you are done. Otherwise, you will continue to be charged!**

**Prerequisites: You need to have Amazon Educate account and a working ssh setup**

Obtain an Amazon AWS account if you have not done it already. Please visit: [AWS Educate](https://aws.amazon.com/edu/)

SSH keypair (one time)
[http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-key-pairs.html#having-ec2-create-your-key-pair](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-key-pairs.html#having-ec2-create-your-key-pair)

**Instructions for quick EMR cluster creation (No Jupyter or Ipython notebook)**

(you can run spark jobs with spark-submit and pyspark)

Spark_EMR_On_AWS.pdf

**Instructions for EMR with Jupyter Notebook (takes at least 30 minutes!)**


1. Login to Amazon AWS console
   [https://aws.amazon.com/emr/](https://aws.amazon.com/emr/)
2. Click on Services at the top left and choose EMR from the panel of services
3. Click on "Create Cluster" and choose "Advanced Options"
4. "Software Configuration": Add Spark to the list and choose "Next" at the bottom
5. "Hardware": You can leave the default hardware (m3.xlarge) and choose the number of core instances. (the default are 1 and 2). You will be charged for this. If you want to know how much you will be charged. visit: [https://aws.amazon.com/emr/pricing/](https://aws.amazon.com/emr/pricing/)
6. General Options:
   a. choose a name
   b. uncheck termination protection
   c. choose Bootstrap Action:
      i. Add Bootstrap Action --> choose "Custom Action"
      --> click "Configure and Add"
      --> Choose any name
      --> Copy and paste the following in "Script Location"
      s3://emr-scripts-kindires/install-jupyter-emr-12212016.sh
7. Security --> if you have ssh keys created (see below), select that for the EC2 keypair. You need this for sshing to the cluster.
8. Create cluster

**Terminating the cluster**

You can terminate the cluster by pressing the terminate button on the cluster "dashboard".

**Working with Spark on the EMR cluster**

Working on the EMR cluster (CS 205)