# CS643 Programming Assignment 2

**John Daudelin**

**12/1/2020**

**CS643-851**

**GitHub Link:** <https://github.com/johndaudelin/cs643-final-project>

**DockerHub Link:**  <https://hub.docker.com/u/johndaudelin98/content> ## CHANGE THIS

**Setup:**

Create an S3 Bucket by navigating in AWS to S3 🡪 Create Bucket. Name the bucket “cs643johndaudelin,” un-select the “Block all public access” box (which may require also selecting the “I acknowledge…” statement that appears, and accept all the other default options and create the bucket. Click on the bucket and hit “Upload.” Then, upload the “TrainingDataset.csv”, “ValidationDataset.csv”, and “TestDataset.csv” files. Since I didn’t have access to the real TestDataset.csv file, I simply copied the ValidationDataset.csv, renamed it to TestDataset.csv, and uploaded this file to the S3 bucket.

Navigate on AWS to EMR 🡪 Create Cluster. Go to Advanced Options. Choose a release of “emr-5.32.0.” For Software Configuration, select Hadoop, Ganglia, Hive, JupyterEnterpriseGateway, Zeppelin, Hue, Spark, and Pig. Don’t change anything else and click Next. Under “Cluster Nodes and Instances,” type 3 for the number of core node instances (in addition to the one master node), delete the Task node group, and keep the default instance type for both the master and core nodes as “m5.xlarge.” Make sure “Cluster scaling” is not selected. Click Next. Name the cluster “CS643 Cluster.” Un-select “Logging” and click Next. Under “Security Options,” select “Proceed without an EC2 key pair.” Hit Create cluster!

This may take a few minutes for the cluster to boot up. Meanwhile, click on “Notebooks” in the left pane, and then hit “Create Notebook.” Name the notebook, and then for “Cluster,” hit “Choose” and select the cluster that you created in the previous step (“CS643 Cluster”). Keep all the other default setting and hit Create notebook. Wait till the status changes from “Starting” to “Ready,” and then click “Open in Jupyter.” Open the notebook that you just created, and then click on Kernel 🡪 Change Kernel 🡪 PySpark on the toolbar at the top.

**Training:**

**Prediction (without Docker):**

Create an EC2 on AWS. SSH into the EC2 and run the following commands to setup the environment:

Install Java.

Install Python 3

Run “pip install pyspark” and “pip install boto3” in the command line.

**Prediction (with Docker):**