CS 643851 – Cloud Computing

Programming Assignment 2

Wine Prediction

This assignment is to develop parallel machine learning (ML) applications in AWS platform. We are using Apache Spark to train an ML model in parallel on multiple EC2 instances. We are also using Spark’s MLlib to develop and use ML model in cloud. Along with that creating a Docker container for the ML model in order to simplify model deployment.

Steps to follow:-

1. First we write and edit the code in Jupyter python notebook and create an EC2 instance in AWS
2. We establish a connection on putty with the newly created instance in order to run our code.
3. Using flintrock, we created a 5 EC2 instances out of which one is master and other 4 are slaves
4. We submit the code saved in our py file from jupyter to master node for parallel execution
5. As an input we put 2 datasets in each of the 4 slave nodes to train our ML model
6. These 2 datasets are TrainingDataSet.csv and ValidationDataset.csv.
7. Training Data set is used to train the model in parallel on multiple EC2 instances
8. Whereas Validation Data Set is used to validate the model and optimize the performance