1. Create AWS EMR Clusters using the key pairs and select the instance as 4(one Master and three slaves) and select application as spark
2. After Creating the Cluster add the SSH inbound rule to the master node for port 22
3. Using putty Connect to the master node using the public ip address and the AWS pem file.
4. We must Install the following packages in the masters in order to run the python code
   1. **pip install pyspark**
   2. **pip install numpy**
5. The Assignment contains two py files
   1. one py file (ModelCreation.py) to read the Trainingdata.csv train and save the model

Command to run the py file: **python ModelCreation.py**

* 1. other py file (Application.py) to read the model and take the test data as input from the command line arguments and gives the F1 score as output

Command with args (Without docker):

**python Applciation.py TestDataset.csv**

1. Create a docker file (DockerFile)

Command: **nano DockerFile**

This file consists of all the commands which are needed to be executed while building the image

1. Install Docker using the command **sudo yum install docker -y**
2. Start the Docker service using **sudo service docker start**
3. Build the image using the command below (which takes DockerFile as input)

**sudo docker build -t assignment2 .**

1. Check whether the image is created or not using the command **sudo docker images**
2. Once the image is created now tag the image to the docker hub.

Command: **sudo docker tag assignment2 vishnureddy119/assignment2**

1. Once the image is tagged now push the image to the docker hub for which we have to login in to the docker. Below are the steps to be followed

**sudo docker login**

**sudo docker push vishnureddy119/assignment2**

1. Once the image is pushed to the Docker hub now, we must pull the image from the docker hub and run it

Commands to run the image as an instance (Container) and execute the application,

**sudo docker run --name model -d assignment2**

copying the testdata.csv into the container

**sudo docker cp ../TestDataset.csv model:/**

get inside the container and execute the application

**sudo docker exec -it model sh**

now running the application to calculate the f1 score

**python Application.py TestDataset.csv**