**Run with Docker**

Create and Select the instance type (e.g., t2.medium).

**Install Java**

sudo yum install -y wget

wget https://corretto.aws/downloads/latest/amazon-corretto-11-x64-linux-jdk.rpm

sudo rpm -ivh amazon-corretto-11-x64-linux-jdk.rpm

**Install and Download Docker**

sudo yum install -y yum-utils device-mapper-persistent-data lvm2

sudo yum-config-manager --add-repo https://download.docker.com/linux/centos/docker-ce.repo

sudo nano /etc/yum.repos.d/docker-ce.repo

**Update the Docker file and Save it.**

[docker-ce-stable]

name=Docker CE Stable - $basearch

baseurl=https://download.docker.com/linux/centos/7/$basearch/stable

enabled=1

gpgcheck=1

gpgkey=https://download.docker.com/linux/centos/gpg

Enter and Save file

**Install Docker container**

sudo yum install -y docker-ce docker-ce-cli containerd.io

sudo systemctl start docker

**Download and Install Apache Spark**

wget https://archive.apache.org/dist/spark/spark-3.1.2/spark-3.1.2-bin-hadoop3.2.tgz

tar xvf spark-3.1.2-bin-hadoop3.2.tgz

sudo mv spark-3.1.2-bin-hadoop3.2 /opt/spark

export SPARK\_HOME=/opt/spark

export PATH=$SPARK\_HOME/bin:$PATH

source ~/.bashrc

**Install pip**

curl https://bootstrap.pypa.io/get-pip.py -o get-pip.py

sudo python3 get-pip.py

pip --version

**Install pyspark**

pip install pyspark

**Install Apache Spark:**

wget https://downloads.apache.org/spark/spark-3.1.2/spark-3.1.2-bin-hadoop3.2.tgz

tar xvf spark-3.1.2-bin-hadoop3.2.tgz

sudo mv spark-3.1.2-bin-hadoop3.2 /opt/spark

echo "export SPARK\_HOME=/opt/spark" >> ~/.bashrc

echo "export PATH=$PATH:$SPARK\_HOME/bin:$SPARK\_HOME/sbin" >> ~/.bashrc

source ~/.bashrc

**Set Java Home to Environment file.**

export JAVA\_HOME=/usr/lib/jvm/java-11-amazon-corretto

export PATH=$JAVA\_HOME/bin:$PATH

/usr/lib/jvm/java-11-amazon-corretto/bin/java

**Install PIP**

sudo yum install python3-pip

pip install numpy

pip install pyspark==3.1.2 > It should be same version as spark-submit –version

pip --version

**Configure Spark for Cluster Mode:**

Edit the spark-env.sh file:

nano $SPARK\_HOME/conf/spark-env.sh

export SPARK\_MASTER\_HOST='your-master-node-ip'

export JAVA\_HOME='/usr/lib/jvm/java-11-amazon-corretto'

**Create log4j.properties file**

sudo touch /home/ec2-user/log4j.properties

**Create Docker file**

touch Dockerfile

nano Dockerfile > Edit file with data then save

sudo docker build -t vivekpa2cs643 . > Command for build image

sudo docker run vivekpa2cs643

sudo docker images > List all Images of Docker

**Run without Docker**

Create and Select the instance type (e.g., t2.medium).

**Install Java**

sudo yum install -y wget

wget https://corretto.aws/downloads/latest/amazon-corretto-11-x64-linux-jdk.rpm

sudo rpm -ivh amazon-corretto-11-x64-linux-jdk.rpm

**Download and Install Apache Spark**

wget https://archive.apache.org/dist/spark/spark-3.1.2/spark-3.1.2-bin-hadoop3.2.tgz

tar xvf spark-3.1.2-bin-hadoop3.2.tgz

sudo mv spark-3.1.2-bin-hadoop3.2 /opt/spark

export SPARK\_HOME=/opt/spark

export PATH=$SPARK\_HOME/bin:$PATH

source ~/.bashrc

**Install pip**

curl https://bootstrap.pypa.io/get-pip.py -o get-pip.py

sudo python3 get-pip.py

pip --version

**Install pyspark**

pip install pyspark

**Install Apache Spark:**

wget https://downloads.apache.org/spark/spark-3.1.2/spark-3.1.2-bin-hadoop3.2.tgz

tar xvf spark-3.1.2-bin-hadoop3.2.tgz

sudo mv spark-3.1.2-bin-hadoop3.2 /opt/spark

echo "export SPARK\_HOME=/opt/spark" >> ~/.bashrc

echo "export PATH=$PATH:$SPARK\_HOME/bin:$SPARK\_HOME/sbin" >> ~/.bashrc

source ~/.bashrc

**Set Java Home to Environment file.**

export JAVA\_HOME=/usr/lib/jvm/java-11-amazon-corretto

export PATH=$JAVA\_HOME/bin:$PATH

/usr/lib/jvm/java-11-amazon-corretto/bin/java

**Install PIP**

sudo yum install python3-pip

pip install numpy

pip install pyspark==3.1.2 > It should be same version as spark-submit –version

pip --version

**Configure Spark for Cluster Mode:**

Edit the spark-env.sh file:

nano $SPARK\_HOME/conf/spark-env.sh

export SPARK\_MASTER\_HOST='your-master-node-ip'

export JAVA\_HOME='/usr/lib/jvm/java-11-amazon-corretto'

**Create log4j.properties file for logs**

sudo touch /home/ec2-user/log4j.properties

**Run Python file**

Python train.py

Python predict.py ValidationDataset.csv

Git hub Repo :- <https://github.com/vivek486/Vivek-pa2-cs643>

Docker Repo :- <https://hub.docker.com/layers/vivekarti/assignment2/v1.0/images/sha256-631318c6273df4546f17ee66fc079beb4b335c9b4fd7c48e8f2c9c1bdfad111e?context=repo>