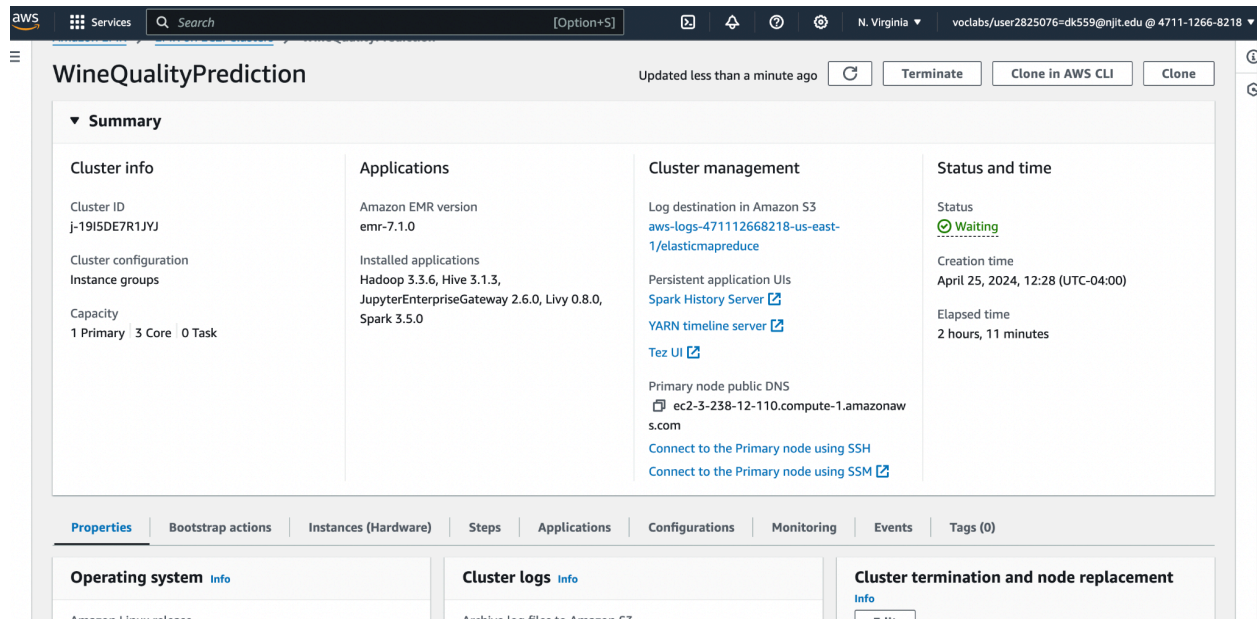


Programming Assignment 2:

In this assignment, EMR cluster with 3 nodes is created.

ec2-3-238-12-110.compute-1.amazonaws.com is the master node and the security group of master node has ssh permissions.



Connect to EC2 Instance - Master node:

1. Open terminal and go to the folder where the key is saved. Change the permission using chmod command.
chmod 400 prediction.pem
2. Later on, connect the ec2 instance with the following command:
ssh -i prediction.pem hadoop@ec2-3-238-12-110.compute-1.amazonaws.com
ec2-3-238-12-110.compute-1.amazonaws.com is the DNS of ec2 instance master node.

Uploaded required files to S3: s3://aws-logs-471112668218-us-east-1/winequality/

```
Last login: Thu Apr 25 10:54:50 on ttys000
complete:13: command not found: compdef
(base) dikshita@Dikshita-Air ~ % cd Downloads
(base) dikshita@Dikshita-Air Downloads % ssh -i prediction.pem hadoop@ec2-3-238-12-110.compute-1.amazonaws.com
The authenticity of host 'ec2-3-238-12-110.compute-1.amazonaws.com (3.238.12.110)' can't be established.
ED25519 key fingerprint is SHA256:T81kHsgeMX7bHbB1/+bX2FJ7P1yYpGa7oUMwrP1QX8.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-3-238-12-110.compute-1.amazonaws.com' (ED25519) to the list of known hosts.
#
##### Amazon Linux 2023
#####
#####
##### https://aws.amazon.com/linux/amazon-linux-2023
#####
Last login: Thu Apr 25 16:34:16 2024

EEEEEEEEEEEEEEEEEEEE MMMMMMMMM MMMMMMMMM RRRRRRRRRRRRRRR
E:::EEEEEEEEEEEEEEEE M:::M::: M:::M::: R:::R:::R:::R:::R
EE:::EEEEEEEEEEEEEEEE M:::M::: M:::M::: R:::RRRRRR:::R
E:::E EEEEE M:::M::: M:::M::: M RRR:::R R:::R
E:::E M:::M::: M:::M::: M:::M::: R:::R R:::R
E:::EEEEEEEEEEEE M:::M M:::M M:::M R:::RRRRRR:::R
E:::EEEEEEEEEEEE M:::M M:::M M:::M R:::RRRRRR:::R
E:::EEEEEEEEEEEE M:::M M:::M M:::M R:::RRRRRR:::R
E:::E M:::M M:::M M:::M R:::R R:::R
E:::E EEEEE M:::M M M:::M R:::R R:::R
EE:::EEEEEEEEEEEE M:::M M:::M R:::R R:::R
E:::EEEEEEEEEEEE M:::M M:::M R:::R R:::R
EEEEEEEEEEEEEEEEEEEE MMMMMMMMM MMMMMMMMM RRRRRRR RRRRRR

[hadoop@ip-172-31-10-247 ~]$ aws s3 cp s3://aws-logs-47112668218-us-east-1/winequality/WineQualityPrediction.py .
download: s3://aws-logs-47112668218-us-east-1/winequality/WineQualityPrediction.py to ./WineQualityPrediction.py
[hadoop@ip-172-31-10-247 ~]$ spark-submit WineQualityPrediction.py
Apr 25, 2024 4:48:11 PM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but /usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location

Traceback (most recent call last):
  File "/home/hadoop/WineQualityPrediction.py", line 1, in <module>
    import numpy as np # linear algebra
ModuleNotFoundError: No module named 'numpy'
24/04/25 16:48:13 INFO ShutdownHookManager: Shutdown hook called
24/04/25 16:48:13 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-2db186ef-5da4-4316-b413-72282c13be2a
[hadoop@ip-172-31-10-247 ~]$ pip install numpy
Defaulting to user installation because normal site-packages is not writeable
Collecting numpy
  Downloading numpy-1.26.4-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (18.2 MB)
    |#####| 18.2 MB 4.9 MB/s
Installing collected packages: numpy
  WARNING: The script f2py is installed in '/home/hadoop/.local/bin' which is not on PATH.
  Consider adding this directory to PATH or, if you prefer to suppress this warning, use --no-warn-script-location.
Successfully installed numpy-1.26.4
[hadoop@ip-172-31-10-247 ~]$ pip install pandas
Defaulting to user installation because normal site-packages is not writeable
Collecting pandas
  Downloading pandas-2.2.2-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (13.1 MB)
    |#####| 13.1 MB 6.4 MB/s
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.9/site-packages (from pandas) (2024.1)
Collecting python-dateutil>=2.8.2
  Downloading python_dateutil-2.9.0.post0-py2.py3-none-any.whl (229 KB)
    |#####| 229 KB 99.1 MB/s
```

aws Services Search [Option+S]

Copy S3 URI

Objects Properties

Objects (6) Info

Copy S3 URI Copy URL Download Open Delete Actions Create folder Upload

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Find objects by prefix

	Name	Type	Last modified	Size	Storage class
	Dockerfile	-	April 25, 2024, 12:18:04 (UTC-04:00)	982.0 B	Standard
	LogisticRegression.model/	Folder	-	-	-
	RandomForestClassifier.model/	Folder	-	-	-
	TrainingDataset.csv	csv	April 25, 2024, 08:18:51 (UTC-04:00)	67.2 KB	Standard
	ValidationDataset.csv	csv	April 25, 2024, 08:18:52 (UTC-04:00)	8.6 KB	Standard
	WineQualityPrediction.py	py	April 25, 2024, 12:18:43 (UTC-04:00)	5.8 KB	Standard

Docker image:

Installed docker on local computer and created docker image and pushed the docker image on > docker hub with the following command.

> docker build --platform linux/amd64 -t wine-quality-docker:v5.0 .

> docker tag wine-quality-docker:v5.0 dk559/wine-quality-prediction:v5.0

> docker push dk559/wine-quality-prediction:v4.0

```
View build details: docker-desktop://dashboard/build/desktop-linux/desktop-linux/mdjimqo58hbzkmzsfv42daf83
(base) dikshita@Dikshitas-Air ProgrammingAssignment % docker build --platform linux/amd64 -t wine-quality-docker:v4.0 .
[+] Building 160.3s (21/21) FINISHED docker:desktop-linux
=> [internal] load build definition from Dockerfile 0.0s
=> => transferring dockerfile: 1.11kB 0.0s
=> [internal] load metadata for docker.io/library/centos 0.9s
=> [auth] library/centos:pull token for registry-1.d 0.0s
=> [internal] load .dockerignore 0.0s
=> => transferring context: 2B 0.0s
=> [ 1/15] FROM docker.io/library/centos:7@sha256:be65f488b7764ad3638f236 4.1s
=> => resolve docker.io/library/centos:7@sha256:be65f488b7764ad3638f236 1.20kB / 1.20kB 0.0s
=> => sha256:be65f488b7764ad3638f236 1.20kB / 1.20kB 0.0s
=> => sha256:dead07b4d8ed7e29e98de0f4504 529B / 529B 0.0s
=> => sha256:eeb6ee3f44bd0b5103bb561 2.75kB / 2.75kB 0.0s
=> => sha256:2d473b07cdd5f0912cd6f 76.10MB / 76.10MB 2.1s
=> => extracting sha256:2d473b07cdd5f0912cd6f1a70335 1.8s
=> [internal] load build context 0.0s
=> => transferring context: 270B 0.0s
=> [ 2/15] RUN yum -y update && yum -y install pyth 68.6s
=> [ 3/15] RUN python -V 0.1s
=> [ 4/15] RUN python3 -V 0.1s
=> [ 5/15] RUN pip3 install --upgrade pip 2.2s
=> [ 6/15] RUN pip3 install numpy panda 5.4s
=> [ 7/15] RUN cd /opt && wget https://apache.osuosl 75.5s
=> [ 8/15] RUN ln -s /opt/spark-3.4.3-bin-hadoop3 /o 0.2s
=> [ 9/15] RUN (echo 'export SPARK_HOME=/opt/spark' 0.1s
=> [10/15] RUN mkdir /wineapp 0.2s
=> [11/15] COPY WineQualityPrediction.py /wineapp/ 0.0s
=> [12/15] RUN rm /bin/sh && ln -s /bin/bash /bin/sh 0.1s
=> [13/15] RUN /bin/bash -c "source ~/.bashrc" 0.2s
=> [14/15] RUN /bin/sh -c "source ~/.bashrc" 0.2s
=> [15/15] WORKDIR /wineapp 0.0s
=> => exporting to image 2.2s
=> => exporting layers 2.2s
=> => writing image sha256:06e6c66ceaf1596021910420b 0.0s
=> => naming to docker.io/library/wine-quality-docke 0.0s

View build details: docker-desktop://dashboard/build/desktop-linux/desktop-linux/tz2aohofc3n36otyurh9tg347

Build multi-platform images faster with Docker Build Cloud: https://docs.docker.com/go/docker-build-cloud
(base) dikshita@Dikshitas-Air ProgrammingAssignment % docker tag wine-quality-docker:v4.0 dk559/wine-quality-prediction:v4.0
(base) dikshita@Dikshitas-Air ProgrammingAssignment % docker push dk559/wine-quality-prediction:v4.0
The push refers to repository [docker.io/dk559/wine-quality-prediction]
5f70bf18a086: Layer already exists
9a9bc32806ed: Pushed
576f14657c96: Pushed
6bb78edbcdf2: Pushed
64e4d4fe1efa: Pushed
b2073b331a8a: Pushed
```

Run the Docker image on EC2 instance:

```
> sudo yum update -y
> sudo yum install docker -y
> sudo service docker start
> sudo usermod -a -G docker ec2-user
> docker run dk559/wine-quality-prediction:v5.0
```

I learned about AWS EMR, Dockerfile and Apache Spark by watching YouTube videos

I had used ChatGPT incase I faced any errors by pasting the errors.

Example: I was facing issue with the Docker run command on EC2 instance after creating the docker image. ChatGPT suggested me to pass the platform parameter while creating the image.



You

The requested image's platform (linux/arm64/v8) does not match the detected host platform (linux/amd64/v3) and no specific platform was requested



ChatGPT

The error message you're encountering indicates a mismatch between the platform of the Docker image you're trying to run and the platform of the host system where Docker is running.

It seems that the Docker image you're trying to run is built for the `linux/arm64/v8` platform, which is for ARM-based systems, while the host system where Docker is running is `linux/amd64`, which is for x86-based systems.

To resolve this issue, you have a few options:

1. **Build Docker Image for the Host Architecture:** If you have access to the Dockerfile used to build the image, you can rebuild the Docker image for the `linux/amd64` platform. You can do this by specifying the `--platform` option when building the image:

bash

Copy code

```
docker build --platform linux/amd64 -t your-image-name .
```



2. **Pull Correct Image for Host Architecture:** If you're pulling the image from a registry, ensure

Option+S

N. Virginia

voclabs/user2825076=dk559@njit.edu @ 4711-1266-8218

Instances (1/7) Info

Find Instance by attribute or tag (case-sensitive)

All states

Launch instances

	Name	Instance ID	Instance state	Instance type	Status check	Alarm status
<input type="checkbox"/>		i-09986f5e70c651fe4	Running	m5.xlarge	2/2 checks passed	View alarms
<input type="checkbox"/>		i-0c2d8b5cf6250fe35	Running	m5.xlarge	2/2 checks passed	View alarms
<input type="checkbox"/>		i-0f65b303135ccad7a	Running	m5.xlarge	2/2 checks passed	View alarms
<input type="checkbox"/>	MasterNode	i-0502ac51b02c8c0ee	Running	m5.xlarge	2/2 checks passed	View alarms
<input checked="" type="checkbox"/>	DockerEC2	i-0a0e3a9d2d0bf09e4	Running	t2.micro	2/2 checks passed	View alarms

i-0a0e3a9d2d0bf09e4 (DockerEC2)

Details

Status and alarms New

Monitoring

Security

Networking

Storage

Tags

▼ Instance summary Info

Instance ID

i-0a0e3a9d2d0bf09e4 (DockerEC2)

IPv6 address

-

Hostname type

IP name: ip-172-31-94-7.ec2.internal

Public IPv4 address

18.234.221.255 | [open address](#)

Instance state

Running

Private IP DNS name (IPv4 only)

ip-172-31-94-7.ec2.internal

Private IPv4 addresses

172.31.94.7

Public IPv4 DNS

ec2-18-234-221-255.compute-1.amazonaws.com | [open address](#)