

Programming Assignment 2

Cloud Computing

Name : Venkanna Dharavath

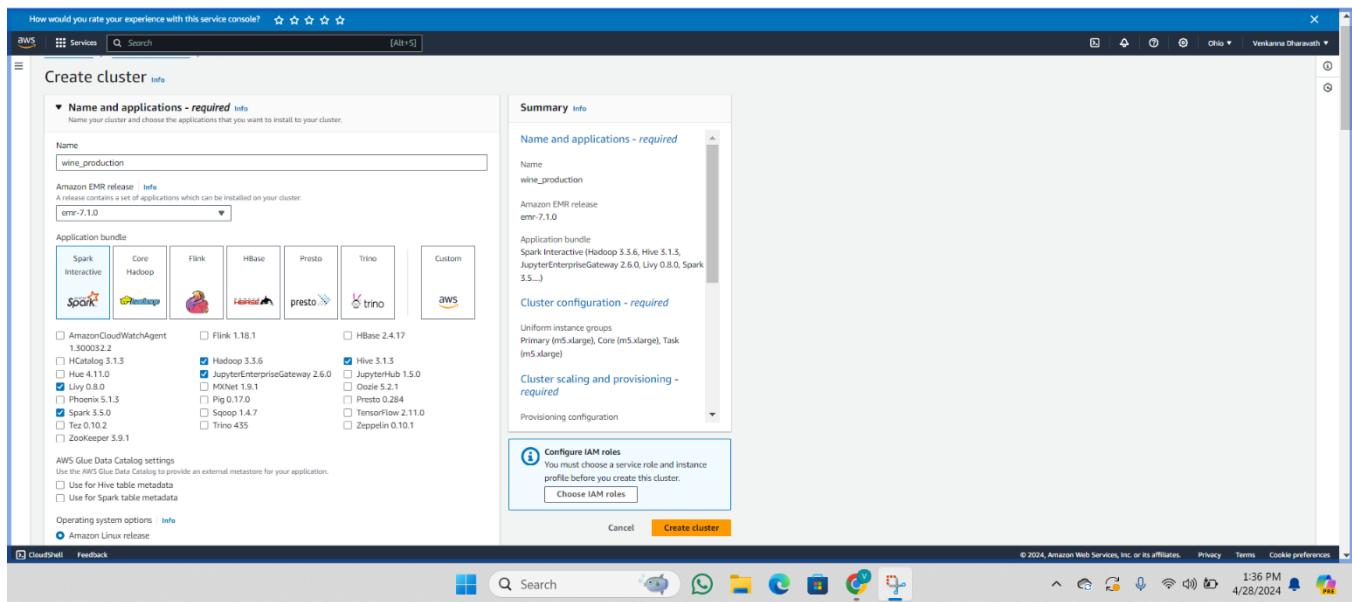
UCID : VD348

GitHub – https://github.com/Venkanna369/wine_production.git

DockerHub - https://hub.docker.com/repository/docker/venkanna369/wine_production/general

1. Log into the AWS Management Console.

The Amazon EMR service, which may be accessed via the service list. Next, select EMR on EC2 Clusters.



2. The Clusters page will open and you will notice that there aren't any clusters that are currently in use.

2. To create a new cluster, click on the "Create cluster" button.

3. You can give the cluster whatever name you like. Additionally, make sure you pick "Spark Interactive" from the Application bundle options and the most recent version of EMR that is available.

1. For the Core, Task, and Primary instance groups, I chose the instance type m5.xlarge. Any instance type that best meets your needs can be chosen by you.

The screenshot shows the AWS EMR Cluster Configuration interface. On the left, there are three sections: Primary, Core, and Task 1 of 1. Each section has a dropdown for 'Choose EC2 instance type' set to 'm5.xlarge'. Below each dropdown is a link to 'Node configuration - optional'. To the right, a summary panel displays the cluster details:

- Name and applications - required**: Name is 'winequality_prediction', Application bundle is 'Spark Interactive (Hadoop 3.3.6, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.8.0, Spark 3.5....)
- Cluster configuration - required**: Uniform instance groups: Primary (m5.xlarge), Core (m5.xlarge), Task (m5.xlarge).
- Cluster scaling and provisioning - required**: Provisioning configuration is set to 'Set cluster size manually'.
- Configure IAM roles**: A note says 'You must choose a service role and instance profile before you create this cluster.' with a 'Choose IAM roles' button.

2. Set the instance sizes of Core and Task-1 to 1 and 2, respectively, to provide cluster scalability and provisioning.

The screenshot shows the 'Cluster scaling and provisioning - required' section of the AWS EMR Cluster Configuration interface. It includes the following options:

- Choose an option**: 'Set cluster size manually' is selected, with a note: 'Use this option if you know your workload patterns in advance.'
- Provisioning configuration**: A table sets the size of core and task instance groups. The table rows are:

Name	Instance type	Instance(s) size	Use Spot purchasing option
Core	m5.xlarge	1	<input type="checkbox"/>
Task - 1	m5.xlarge	2	<input type="checkbox"/>

To the right, the summary panel shows the updated cluster details:

- Name and applications - required**: Name is 'wine_production', Application bundle is 'Spark Interactive (Hadoop 3.3.6, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.8.0, Spark 3.5....)
- Cluster configuration - required**: Uniform instance groups: Primary (m5.xlarge), Core (m5.xlarge), Task (m5.xlarge).
- Cluster scaling and provisioning - required**: Provisioning configuration is now set to 'Use EMR-managed scaling'.

3. In order to pick the security groups for EC2's Primary node, Core, and task nodes, select the security groups that are displayed below.

The screenshot shows the AWS EMR Cluster Configuration page. Under 'Cluster termination and node replacement', the 'Termination option' is set to 'Manually terminate cluster'. The 'Use termination protection' checkbox is checked. Under 'Unhealthy node replacement - new', the 'Turn on' radio button is selected. The 'Bootstrap actions' section is empty. The 'Cluster logs' section is collapsed. On the right side, the 'Summary' tab is selected, showing the cluster name 'wine_production' and application bundle 'Amazon EMR release emr-7.1.0'. Below that are sections for 'Cluster configuration - required' (Uniform instance groups: Primary (m5.xlarge), Core (m5.xlarge), Task (m5.xlarge)) and 'Cluster scaling and provisioning - required' (Provisioning configuration).

4. To prevent the cluster from being terminated automatically, make sure that the option to manually end the cluster is selected.

This screenshot is identical to the one above, showing the AWS EMR Cluster Configuration page. The 'Termination option' remains set to 'Manually terminate cluster'. The 'Use termination protection' checkbox is checked. Under 'Unhealthy node replacement - new', the 'Turn on' radio button is selected. The 'Bootstrap actions' section is empty. The 'Cluster logs' section is collapsed. The right-hand sidebar shows the 'Summary' tab selected, displaying the cluster name 'wine_production', application bundle 'Amazon EMR release emr-7.1.0', and the required instance groups and provisioning configuration.

5.

6. create a new key pair

The screenshot shows the 'Security configuration and EC2 key pair' section. It includes fields for choosing a security configuration, creating a new one, and selecting an EC2 key pair. A note says 'You haven't entered an EC2 key. If you're outside a VPN and want to enable SSH or use Hue SQL assistant with this cluster, you must enter an EC2 key.' The 'Identity and Access Management (IAM) roles - required' section shows options for choosing an existing service role or creating a new one. The 'Amazon EMR service role' section indicates that a service role is required for provisioning resources. The 'Summary' section on the right shows the cluster name 'wine_production' and its Amazon EMR release 'emr-7.1.0'. The 'Cluster configuration - required' section lists uniform instance groups: Primary (m5.xlarge), Core (m5.xlarge), Task (m5.xlarge). The 'Cluster scaling and provisioning - required' section shows a provisioning configuration.

7. Provide a custom name for the key pair and select 'ppk' as the file format when creating the key pair

The screenshot shows the 'Create key pair' page. It has fields for 'Name' (set to 'key_pair'), 'Key pair type' (set to 'RSA'), and 'Private key file format' (set to '.pem'). There are also sections for 'Tags - optional' and 'Add new tag'. At the bottom, there are 'Cancel' and 'Create key pair' buttons.

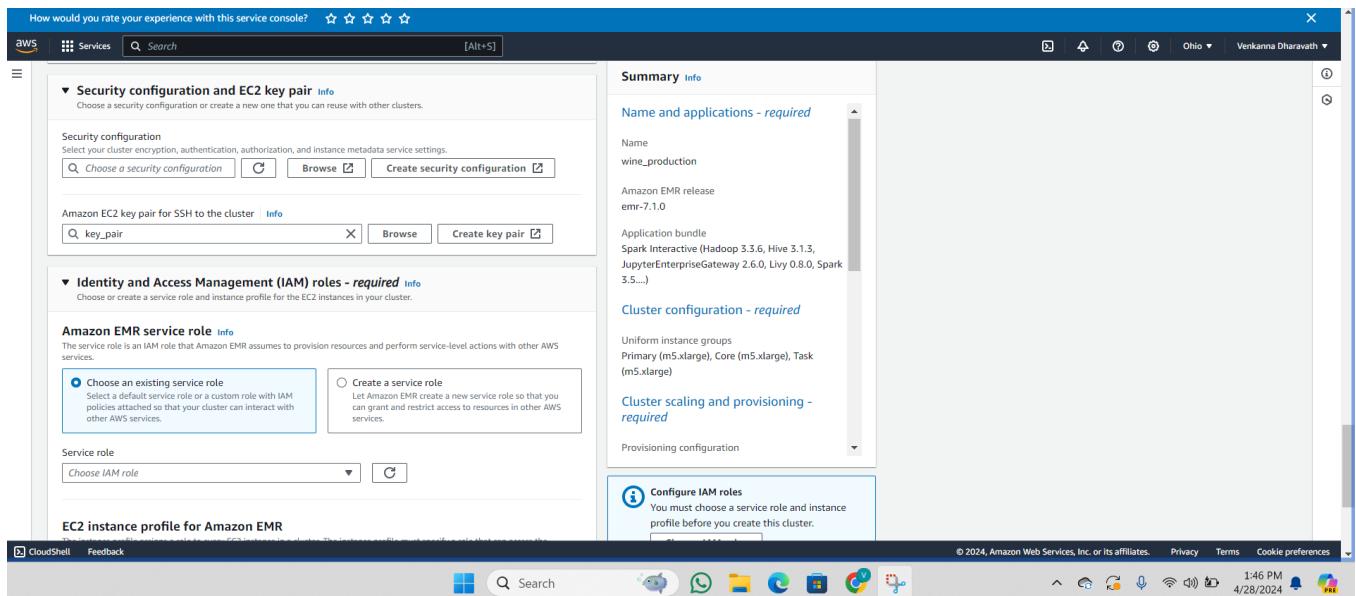
8. The key pair named 'programming-2' has been created, and the key has been downloaded and stored on the local system for connecting to the cluster using PuTTY as the SSH server.

The screenshot shows the AWS EC2 Key Pairs page. The left sidebar includes options like EC2 Dashboard, EC2 Global View, Events, Instances (with sub-options like Instances, Instance Types, Launch Templates, Spot Requests, Savings Plans, Reserved Instances, Dedicated Hosts, Capacity, Reservations), Images (AMIs, AMI Catalog), Elastic Block Store (Volumes, Snapshots, Lifecycle Manager), and Network & Security (Security Groups). The main content area displays a table titled 'Key pairs (1) Info'. The table has columns for Name, Type, Created, Fingerprint, and ID. One row is listed: 'key_pair' (rsa, 2024/04/28 13:45 GMT-4, ffeef9c1c49:65:b1:3c:2e:6b:ee:97:f3:eb:c4:e1:cf:3..., key-00043409fb3db0db7). There are 'Actions' and 'Create key pair' buttons at the top right of the table.

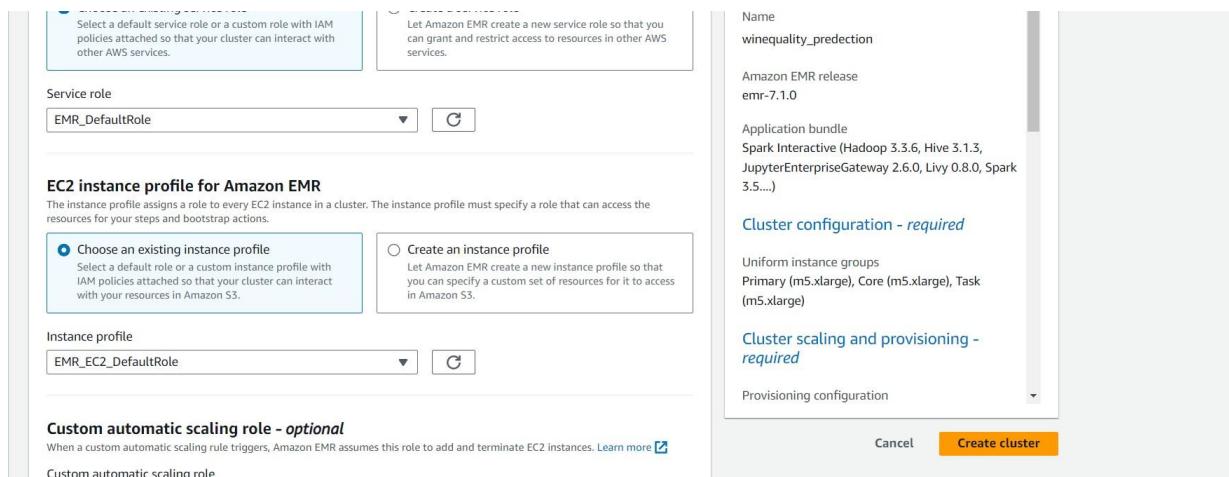
9. The key pair named 'programming-2' has been created, and you can now browse the key .

This screenshot is identical to the one above, showing the AWS EC2 Key Pairs page with a single key pair named 'key_pair'. The interface and data are the same, including the table structure and the 'Create key pair' button.

10. Go to the IAM roles and choose them as indicated below.



11. Click the button labeled "Create Cluster" to initiate the process of creating the clusters.



12. Cluster created successfully.

13. Navigate to the EC2 instance page. As seen below, there are four EC2 instances launched, one of which is a Master node and the other three are Slave nodes.

Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone	Public IPv4 DNS
i-0f7ca45c3a5f4dd63	Running	m5.xlarge	Initializing	View alarms	us-east-1f	ec2-44-211-27-0
i-0e63c06c8bc2578bf	Running	m5.xlarge	Initializing	View alarms	us-east-1f	ec2-35-170-74-0
i-07723f20b4e5321f4	Running	m5.xlarge	Initializing	View alarms	us-east-1f	ec2-18-232-153-
i-0c1a32e92b5b911bc	Running	m5.xlarge	Initializing	View alarms	us-east-1f	ec2-3-235-79-90

Elastic IP	IPv6 IPs	Monitoring	Security group name	Key name	Launch time	Platform
-	-	disabled	ElasticMapReduce-master	programming-2	2024/04/27 13:43 GMT-4	Linux/UNIX
-	-	disabled	ElasticMapReduce-slave	programming-2	2024/04/27 13:43 GMT-4	Linux/UNIX
-	-	disabled	ElasticMapReduce-slave	programming-2	2024/04/27 13:43 GMT-4	Linux/UNIX
-	-	disabled	ElasticMapReduce-slave	programming-2	2024/04/27 13:43 GMT-4	Linux/UNIX

14. To access the "ElasticMapReduce-Master" security group, navigate to the EC2 service and select the appropriate Security ID.

EC2 > Instances > i-086c27f6e4a44c774

Instance summary for i-086c27f6e4a44c774

Updated less than a minute ago

Instance ID	i-086c27f6e4a44c774	Public IPv4 address	18.217.161.57 open address
IPv6 address	-	Instance state	Running
Hostname type	IP name: ip-172-31-47-91.us-east-2.compute.internal	Private IP DNS name (IPv4 only)	ip-172-31-47-91.us-east-2.compute.internal
Answer private resource DNS name	-	Instance type	m5.xlarge
Auto-assigned IP address	18.217.161.57 [Public IP]	VPC ID	vpc-0d3b00305a5dd1206
IAM Role	AmazonEMR-InstanceProfile-20240428T141425	Subnet ID	subnet-01959d18735d1c28c
IMDSv2	Required		

Details Status and alarms New Monitoring Security Networking Storage Tags

CloudShell Feedback © 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences 2:20 PM 4/28/2024

15.

16. Click on the security group to edit its inbound rules

EC2 Dashboard Services Search [Alt+S]

EC2 Global View Events

Instances Instances Types Launch Templates Spot Requests Savings Plans Reserved Instances Dedicated Hosts Capacity Reservations New

Images AMIs AMI Catalog

Elastic Block Store Volumes Snapshots Lifecycle Manager

Network & Security Security Groups

Details Status and alarms New Monitoring Security Networking Storage Tags

Security details

IAM Role AmazonEMR-InstanceProfile-20240428T141425 Owner ID 637423443387 Launch time Sun Apr 28 2024 14:15:22 GMT-0400 (Eastern Daylight Time)

Inbound rules

Name	Security group rule ID	Port range	Protocol	Source	Security groups	Description
-	sgr-05721dec71d64d178	All	ICMP	sg-083cd2bdea936dbf8	ElasticMapReduce-master	-
-	sgr-082926c0d1935f241	All	ICMP	sg-0b52e9b0ae2b46bae	ElasticMapReduce-master	-
-	sgr-02d614607feaf14df	0 - 65535	UDP	sg-0b52e9b0ae2b46bae	ElasticMapReduce-master	-
-	sgr-0e445598d02f844ce	0 - 65535	TCP	sg-0b52e9b0ae2b46bae	ElasticMapReduce-master	-
-	sgr-0c59ff6f50d14d87d	0 - 65535	TCP	sg-083cd2bdea936dbf8	ElasticMapReduce-master	-
-	sgr-0a3c6688e6680caf4	8443	TCP	pl-eca74285	ElasticMapReduce-master	-
-	sgr-016f1e235fb618ac	0 - 65535	UDP	sg-083cd2bdea936dbf8	ElasticMapReduce-master	-

Outbound rules

CloudShell Feedback © 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences 2:20 PM 4/28/2024

17. In the Inbound Rules section, select Edit Inbound Rules.

The screenshot shows the AWS EC2 Security Groups interface. On the left, a sidebar lists various services like EC2 Dashboard, Instances, Images, and Network & Security. The main content area shows a success message: "Inbound security group rules successfully modified on security group sg-083cd2bdea936dbf8 | ElasticMapReduce-master". Below this, the security group name is "sg-083cd2bdea936dbf8 - ElasticMapReduce-master". The "Details" tab is selected, displaying information such as the security group name, ID, owner, and rule counts. The "Inbound rules" tab is active, showing a table with 9 entries. The table columns include Name, Security group rule..., IP version, Type, Protocol, Port range, Source, and Description. A search bar and filter options are at the top of the table. The bottom of the screen shows the AWS navigation bar with links for CloudShell, Feedback, and various icons.

The screenshot shows the AWS Security Groups interface. At the top, there's a search bar and filter options for services like All TCP, TCP, and Custom. Below the search bar, several security groups are listed with their respective rules:

- sgr-0e415598602f844c0: All TCP, TCP, 0 - 65535, Custom, sg-0b52e9b0ae2b46bae
- sgr-0c59ff6f50d14d87d: All TCP, TCP, 0 - 65535, Custom, sg-083cd2bdea936dbf8
- sgr-0a3c6688e6680caf4: Custom TCP, TCP, 8443, Custom, pl-eca74285
- sgr-016f1e235fb618ac: All UDP, UDP, 0 - 65535, Custom, sg-083cd2bdea936dbf8
- : SSH, TCP, 22, Anywhere..., 0.0.0.0/0
- : Custom TCP, TCP, 4040, Anywhere..., 0.0.0.0/0

At the bottom left is an "Add rule" button. A warning message at the bottom states: "⚠ Rules with source of 0.0.0.0/0 or ::/0 allow all IP addresses to access your instance. We recommend setting security group rules to allow access from known IP addresses only." The browser toolbar at the bottom includes CloudShell, Feedback, and various icons.

18. With the parameters listed below, enter the port numbers 22 and 4040. Click the Add rule button to save the rules.

This screenshot shows the AWS Security Groups interface for a specific security group. The rules listed are:

- sgr-0a8de41c5t5zta45: All ICMP - IPv4, ICMP, All, Custom, sg-0f3d5c33b4411ee5e
- : SSH, TCP, 22, My IP, allowing access from my home, 73.112.127.74/32
- : Custom TCP, TCP, 4040, Custom, 0.0.0.0/16, for web interface of spark, 0.0.0.0/16

At the bottom left is an "Add rule" button. A warning message at the bottom states: "⚠ Rules with source of 0.0.0.0/0 or ::/0 allow all IP addresses to access your instance. We recommend setting security group rules to allow access from known IP addresses only." The browser toolbar at the bottom includes CloudShell, Feedback, and various icons.

The screenshot shows the AWS EC2 Security Groups page. A modal window is open at the top left, indicating that inbound security group rules were successfully modified on security group sg-083cd2bdea936dbf8 | ElasticMapReduce-master. The main content area displays the details for this security group. Key information includes:

- Security group name:** ElasticMapReduce-master
- Security group ID:** sg-083cd2bdea936dbf8
- Description:** Master group for Elastic MapReduce created on 2024-04-20T22:40:17.189Z
- VPC ID:** VPC-0d3b09305a5dd1206
- Owner:** 637423443387
- Inbound rules count:** 9 Permission entries
- Outbound rules count:** 1 Permission entry

The interface includes tabs for Inbound rules, Outbound rules, and Tags. Below the main content, there is a table for Inbound rules with columns for Name, Security group rule..., IP version, Type, Protocol, Port range, Source, and Description. A search bar and filter options are also present.

18. Create an S3 bucket in AWS services to store the dataset.

19. Choose "Create Bucket".

The screenshot shows the AWS S3 Buckets page. On the left, a sidebar lists various S3 features like Buckets, Access Grants, and Storage Lens. The main content area shows the General purpose buckets section. It displays a single bucket named "aws-logs-637423443387-us-east-2" located in the US East (Ohio) region. The bucket was created on April 28, 2024, at 14:14:42 (UTC-04:00). The page includes a search bar, filter options for Name, AWS Region, IAM Access Analyzer, and Creation date, and buttons for Copy ARN, Empty, Delete, and Create bucket.

20.

21. Label your bucket as "dataset-programming-assignment-2". Scroll down and click on the "Create bucket" button.

The screenshot shows the AWS S3 console interface. On the left, there is a navigation sidebar with the following sections:

- Buckets
- Access Grants
- Access Points
- Object Lambda Access Points
- Multi-Region Access Points
- Batch Operations
- IAM Access Analyzer for S3

Below these, under "Storage Lens", are:

- Dashboards
- Storage Lens groups
- AWS Organizations settings

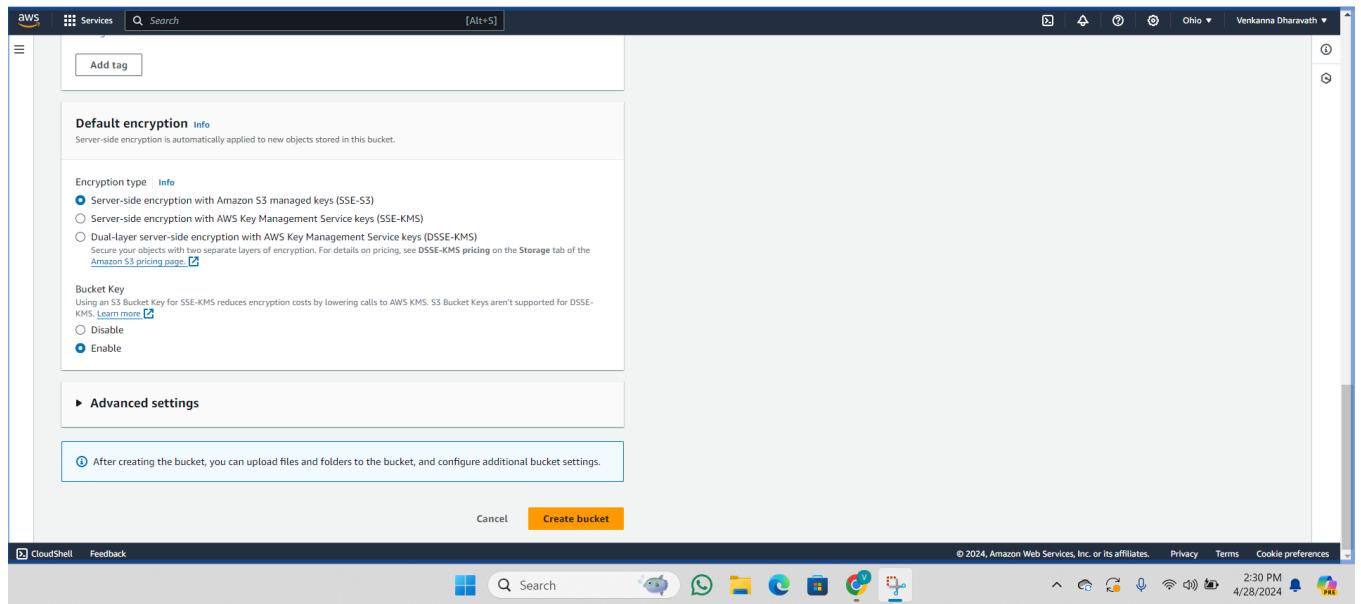
At the bottom of the sidebar are links for "Feature spotlight" (with 7 items) and "AWS Marketplace for S3".

The main content area is titled "Amazon S3" and shows an "Account snapshot" with a link to "All AWS Regions". It includes a "View Storage Lens dashboard" button. Below this, there are two tabs: "General purpose buckets" (selected) and "Directory buckets".

The "General purpose buckets" section has a header "General purpose buckets (1) Info All AWS Regions" and a note that buckets are containers for data stored in S3. It features a search bar "Find buckets by name" and a table with the following data:

Name	AWS Region	IAM Access Analyzer	Creation date
aws-logs-657423445387-us-east-2	US East (Ohio) us-east-2	View analyzer for us-east-2	April 28, 2024, 14:14:42 (UTC-04:00)

At the top right of the main content area are buttons for "Create bucket" (orange), "Copy ARN", "Empty", and "Delete". At the bottom right are navigation icons and the date/time "4/28/2024 2:26 PM".



22. Access the buckets page to view your newly created bucket.

23. Click on the bucket name you created.

The screenshot shows the AWS S3 console. At the top, a green banner displays the message: "Successfully created bucket 'dataset-programming-assignment2'. To upload files and folders, or to configure additional bucket settings, choose View details." Below the banner, the navigation bar shows "Amazon S3 > Buckets". A "Account snapshot" section is present, along with a link to "View Storage Lens dashboard". The main area lists "General purpose buckets (2)" under "All AWS Regions". The table shows two buckets: "aws-logs-657423443387-us-east-2" and "dataset-programming-assignment2". Both buckets are located in "US East (Ohio) us-east-2". The "Creation date" for both is "April 28, 2024, 14:14:42 (UTC-04:00)". Action buttons for each row include "Copy ARN", "Empty", "Delete", and "Create bucket". The bottom of the screen includes standard AWS navigation links like CloudShell, Feedback, and a footer with copyright information.

The screenshot shows the "dataset-programming-assignment2" bucket's objects page. The navigation bar shows "Amazon S3 > Buckets > dataset-programming-assignment2". The main content area is titled "dataset-programming-assignment2" with a "Info" link. Below it is a "Objects (0) Info" section. It states: "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." A "Find objects by prefix" search bar is available. A table header for objects includes columns: Name, Type, Last modified, Size, and Storage class. A message below the table says: "No objects. You don't have any objects in this bucket." An "Upload" button is located at the bottom of the table area. The bottom of the screen includes standard AWS navigation links like CloudShell, Feedback, and a footer with copyright information.

24. Click the upload button.

25.

27. From your local system, choose the.csv files by selecting "Add Files." Now, to upload the dataset to the S3 bucket, click the "Upload button". Your S3 bucket now contains the ValidationDataset.csv and TrainingDataset.csv files.

Upload: status

The information below will no longer be available after you navigate away from this page.

Summary

Destination	Succeeded	Failed
s3://dataset-programming-assignment2	2 files, 75.7 KB (100.00%)	0 files, 0 B (0%)

Files and folders (2 Total, 75.7 KB)

Name	Folder	Type	Size	Status	Error
TrainingDat...	-	text/csv	67.2 KB	Succeeded	-
ValidationD...	-	text/csv	8.6 KB	Succeeded	-

26. Connect to the server using PuTTY by specifying the PPK file for authentication.

Instance summary for i-0f7ca45c3a5f4dd63

Updated less than a minute ago

Actions

Putty Configuration

Session

- Host Name (or IP address): 44.211.27.0
- Port: 22
- Connection type: SSH
- Saved Sessions: Default Settings
- Close window on exit: Only on clean exit

Private IPv4 addresses

- 172.31.76.1

Public IPv4 DNS

- ec2-44-211-27-0.compute-1.amazonaws.com

Elastic IP addresses

-

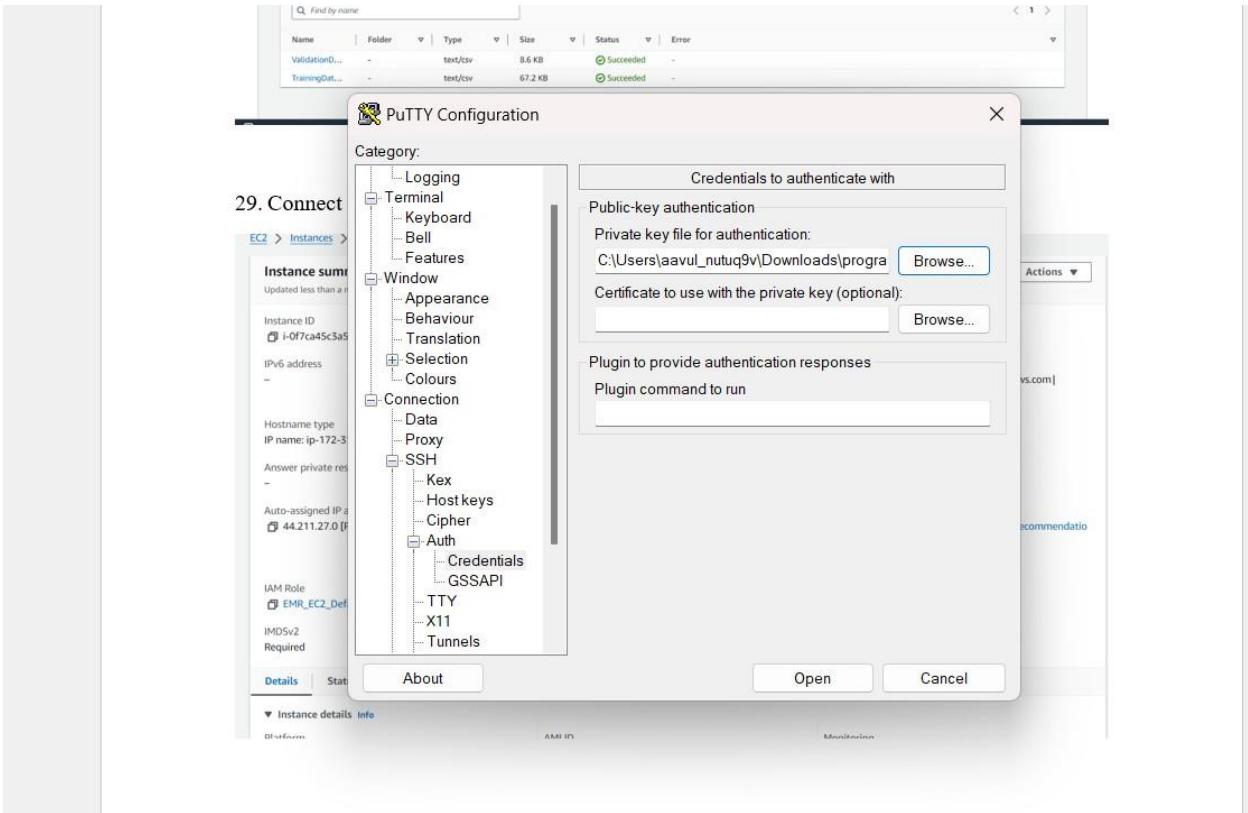
AWS Compute Optimizer finding

- Opt-in to AWS Compute Optimizer for recommendations.

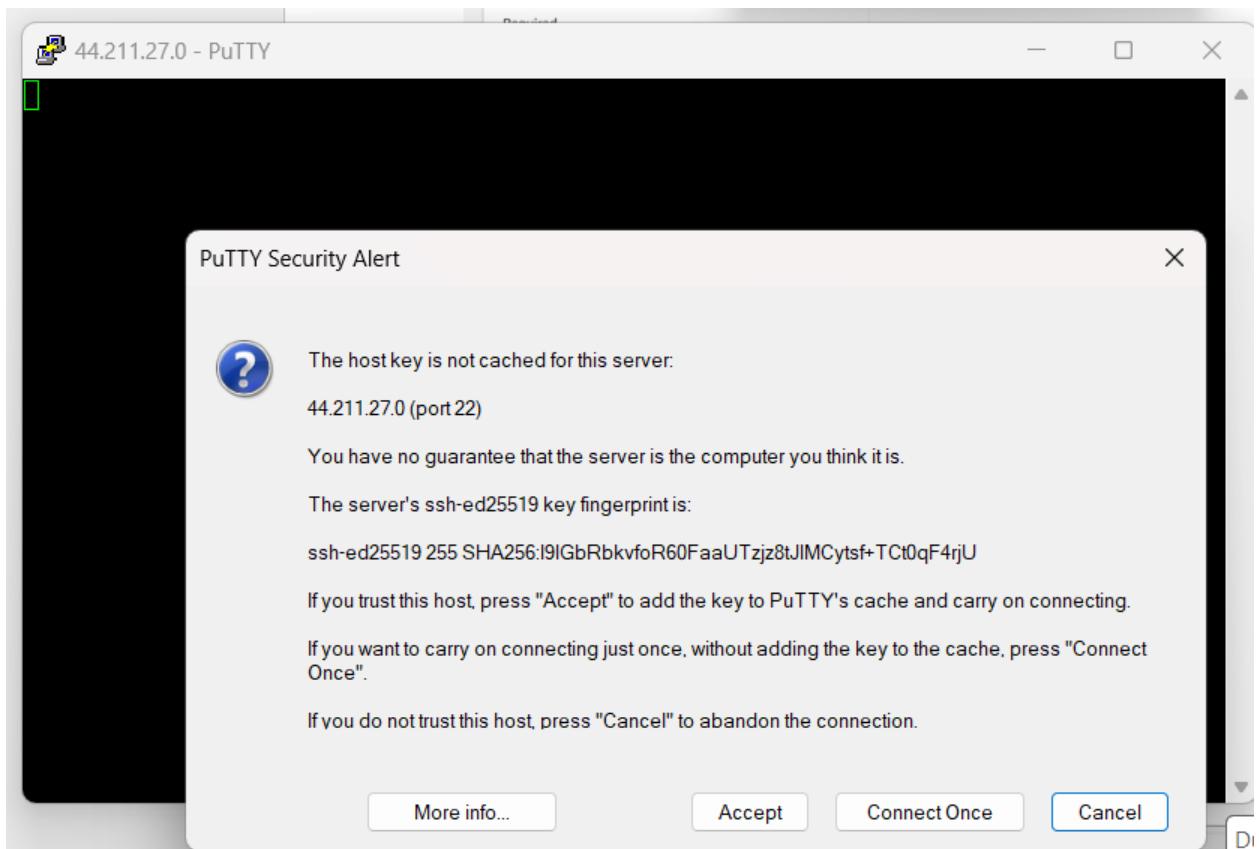
Auto Scaling Group name

-

27. Click on "SSH" under "Auth", select "Credentials", provide the path to the location of the PPK file, and then click "Open".



28. click on accept



29. provide ec2-user

```
ec2-user@ip-172-31-76-1:~$ login as: ec2-user
Authenticating with public key "programming-2"
'          #
~\ _###_      Amazon Linux 2023
~~ \##|
~~ \#/   https://aws.amazon.com/linux/amazon-linux-2023
~~ V~, _->
~~ ./
~~ ./. /_
~~ ./. /_
~/m/,'

EEEEEEEEEEEEEEEEEE MMMMMMM MBBBBBBB RRRRRRRRRRRRRR
E:::::E:::::E:::::E M:::::M M:::::M R:::::R R:::::R
EE::::EEEEE::::E M:::::M M:::::M R:::::R R:::::R
E::::E     EEEEE M:::::M M:::::M RR:::::R R:::::R
E::::E     M:::::M:::M M:::::M:::M R:::::R R:::::R
E:::::EEEEE::::E M:::::M M:::::M M:::::M R:::::R R:::::R
E:::::::::::E M:::::M M:::::M M:::::M R:::::R R:::::R
E:::::EEEEE::::E M:::::M M:::::M M:::::M R:::::R R:::::R
E:::::E     M:::::M M:::::M M:::::M R:::::R R:::::R
E:::::E     EEEE M:::::M MMM M:::::M R:::::R R:::::R
EE::::EEEEE::::E M:::::M M:::::M R:::::R R:::::R
E:::::::::::E M:::::M M:::::M RR:::::R R:::::R
EEEEEEEEEEEEEEEEEE MMMMMMM MBBBBBBB RRRRRRRR

[ec2-user@ip-172-31-76-1 ~]$
```

32. Configure the credentials in the EC2 instance of the Master node.

Enter these commands into your terminal to set up the master node's credentials:

```
# mkdir .aws
# touch .aws/credentials# vi
.aws/credentials
```

30. Copy and paste the credentials from the AWS Academy page and AWS information.

AWS CLI:
Copy and paste the following into
.aws/credentials

```
[default]
aws_access_key_id=ASIAZQ3DNPZB6BB40R3V
aws_secret_access_key=tBakn87qp0aG3DEZuQbwA
ztYyqE+nsxSKNg05CmM
aws_session_token=IQoJb3JpZ21uX2VjEJ
D//////////wEaCXVzLXd1c3QtMiJHMEUCIQD1LrDt1
2Pe7X+yq4qocuPTKZrk99LhhjirjsYdc2KEcQIgYR+8
jKUnJ5LkUYHJmRI8hsX1Ohbhsoij8IMIw8bvo78qsQI
I2f//////////ARAAAGgw2NTQ2NTQxNDQwNjciDJi0Dc
EsoZPX4UUUiRyqFAjDLpepf0M1tZKH+2g6My29KtGSPY
Hfm1Pmd091c00XbhQqyH4tt0XdJchxjF9qjDaYKji+r
Ts5pIX4IHGXGygXHDt3wG6H91dafwkoZsoTj7L1XcDWG
n4FERPjmuZnsZcWT2JMu+eLpMkhgDOs+TKdpey/1or6
PzrykOIBFCbBBEdbCXzga6JJkW/biKTqDkJsxOxc54S
Zr0Cy6KGRcVlpq7mm6qimMZryT0uUMJX9n+roTDn/81
RLhLwaexFsQqX01WEh10KMLPOB8R1mk0TYt316S3GT/
Y+X/mM5aT/Tjg5H/n20+uGoEiW0nGrgBzMvzqypu8PF
1DSJ6k6RDoc2SxYAPfuDCDv7SxBjqdAbIUCws9is91W
wuV5hbII7xXjr14/wgOEc9X+R4G8WWtZ7pPQmKqc5FL
DkXA9qCnp0R1DXcIIi04EBwB4cucgzXyFRXFYvznpEODVaMrw4ayhDQ
3xLF2qMeZ+A9OLGOzUBj+U67MsLDbWSKMzQod/L1bmC
```

31.To execute our Spark application, we must first install the required packages.

Run the commands below:

```
# sudo yum update.
# sudo yum install git.
# pip install pyspark findspark boto3 numpy pandas scikit-learn datetime
```

32.To clone the GitHub repository, enter the following command:

```
#git clone https://github.com/gayatriaavula/winepred_quality.git
```

33.To get started with the Spark Application, perform the following commands:

```
# spark-submit --master yarn CS643_Programming_assignment_2/WineTraining.py # spark-
submit --master yarn CS643_Programming_assignment_2/WineTesting.py >output.txt
```

```
[hadoop@ip-172-31-69-133 ~]$ spark-submit --master yarn CS643_Programming_assignment_2/WineTraining.py
Apr 26, 2024 8:56:28 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
24/04/26 20:56:32 INFO SparkContext: Running Spark version 3.5.0-amzn-1
24/04/26 20:56:32 INFO SparkContext: OS info Linux, 6.1.84-99.169.amzn2023.x86_64, amd64
24/04/26 20:56:32 INFO SparkContext: Java version 17.0.10
24/04/26 20:56:32 INFO ResourceUtils: =====
24/04/26 20:56:32 INFO ResourceUtils: No custom resources configured for spark.driver.
24/04/26 20:56:32 INFO ResourceUtils: =====
24/04/26 20:56:32 INFO SparkConf: Submitted application: WineQuality Training
24/04/26 20:56:32 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 4, script: , vendor: , memory -> name: memory, amount: 9486, s
cript: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ) task resources: Map(cpu -> name: cpus, amount: 1.0)
24/04/26 20:56:32 INFO ResourceProfile: Limiting resource is cpus at 4 tasks per executor
24/04/26 20:56:32 INFO ResourceProfileManager: Added ResourceProfile id: 0
24/04/26 20:56:32 INFO SecurityManager: Changing view acls to: hadoop
24/04/26 20:56:32 INFO SecurityManager: Changing modify acls to: hadoop
24/04/26 20:56:32 INFO SecurityManager: Changing view acls groups to:
24/04/26 20:56:32 INFO SecurityManager: Changing modify acls groups to:
24/04/26 20:56:32 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: hadoop; groups with view permissions: EMPTY; users with modi
ty permissions: hadoop; groups with modify permission: EMPTY
24/04/26 20:56:32 INFO Utils: Successfully started service 'sparkDriver' on port 40123.
24/04/26 20:56:32 INFO SparkEnv: Registering OutputTracker
24/04/26 20:56:32 INFO SparkEnv: Registering BlockManagerMaster
24/04/26 20:56:32 INFO BlockManagerMasterEndpoint: Using Apache spark.storage.DefaultTopologyMapper for getting topology information
24/04/26 20:56:32 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
24/04/26 20:56:32 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
24/04/26 20:56:32 INFO DiskBlockManager: Created local directory at /mnt/tmp/blockmgr-19a6d5b6-a51f-4d1b-b115-eb9b9f2239af
24/04/26 20:56:32 INFO MemoryStore: MemoryStore started with capacity 1048.6 Mib
24/04/26 20:56:32 INFO SparkEnv: Registering OutputCommitCoordinator
24/04/26 20:56:32 INFO SubResultCacheManager: Sub-result caches are disabled.
24/04/26 20:56:32 INFO JettyUtils: Start Jetty 0.0.0.0:4040 for SparkUI
24/04/26 20:56:33 INFO Utils: Successfully started service 'SparkUI' on port 4040.
24/04/26 20:56:33 INFO Utils: Using 50 preallocated executors [minExecutors: 0]. Set spark.dynamicAllocation.preallocateExecutors to 'false' disable executor preallocation.
24/04/26 20:56:33 INFO DefaultNoHARMfollowerProxyProvider: Connecting to ResourceManager at ip-172-31-69-133.ec2.internal/172.31.69.133:8032
24/04/26 20:56:33 INFO Configuration: resource-types.xml not found
24/04/26 20:56:33 INFO ResourceUtils: Unable to find 'resource-types.xml'.
24/04/26 20:56:33 INFO Client: Verifying our application has not requested more than the maximum memory capability of the cluster (12288 MB per container)
24/04/26 20:56:33 INFO Client: Will allocate AM container, with 896 MB memory including 384 MB overhead
24/04/26 20:56:33 INFO Client: Setting up container launch context for our AM
24/04/26 20:56:33 INFO Client: Setting up the launch environment for our AM container
24/04/26 20:56:33 INFO Client: Preparing resources for our AM container
24/04/26 20:56:33 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK HOME.
24/04/26 20:56:34 INFO Client: Uploading resource file:/mnt/tmp/spark-4db2f6a-0ddf-4955-bde0-c52a65fe4c83/_spark_libs_1278197152423226073.zip -> hdfs://ip-172-31-69-133.ec2.internal:8020
24/04/26 20:56:35 INFO Client: Uploading resource file:/etc/spark/conf.dist/hive-site.xml -> hdfs://ip-172-31-69-133.ec2.internal:8020/user/hadoop/.sparkStaging/application_1714161677101_00
03/hive-site.xml
24/04/26 20:56:35 INFO Client: Uploading resource file:/etc/hudi/conf.dist/hudi-defaults.conf -> hdfs://ip-172-31-69-133.ec2.internal:8020/user/hadoop/.sparkStaging/application_171416167710
1_0003/hudi-defaults.conf
24/04/26 20:56:35 INFO Client: Uploading resource file:/usr/lib/spark/python/lib/pyspark.zip -> hdfs://ip-172-31-69-133.ec2.internal:8020/user/hadoop/.sparkStaging/application_1714161677101
_0003/pyspark.zip
24/04/26 20:56:35 INFO Client: Uploading resource file:/usr/lib/spark/python/lib/py4j-0.10.9.7-src.zip -> hdfs://ip-172-31-69-133.ec2.internal:8020/user/hadoop/.sparkStaging/application_171
4161677101_0003/py4j-0.10.9.7-src.zip
```

```
24/04/26 20:56:35 INFO Utils: Successfully started service 'SparkUI' on port 4040.
24/04/26 20:56:35 INFO Utils: Using 50 preallocated executors [minExecutors: 0]. Set spark.dynamicAllocation.preallocateExecutors to 'false' disable executor preallocation.
24/04/26 20:56:35 INFO Configuration: resource-types.xml not found
24/04/26 20:56:35 INFO ResourceUtils: Unable to find 'resource-types.xml'.
24/04/26 20:56:35 INFO Client: Verifying our application has not requested more than the maximum memory capability of the cluster (12288 MB per container)
24/04/26 20:56:35 INFO Client: Will allocate AM container, with 896 MB memory including 384 MB overhead
24/04/26 20:56:35 INFO Client: Setting up container launch context for our AM
24/04/26 20:56:35 INFO Client: Setting up the launch environment for our AM container
24/04/26 20:56:35 INFO Client: Preparing resources for our AM container
24/04/26 20:56:35 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK HOME.
24/04/26 20:56:34 INFO Client: Uploading resource file:/mnt/tmp/spark-4db2f6a-0ddf-4955-bde0-c52a65fe4c83/_spark_libs_1278197152423226073.zip -> hdfs://ip-172-31-69-133.ec2.internal:8020
24/04/26 20:56:35 INFO Client: Uploading resource file:/etc/spark/conf.dist/hive-site.xml -> hdfs://ip-172-31-69-133.ec2.internal:8020/user/hadoop/.sparkStaging/application_1714161677101_00
03/hive-site.xml
24/04/26 20:56:35 INFO Client: Uploading resource file:/etc/hudi/conf.dist/hudi-defaults.conf -> hdfs://ip-172-31-69-133.ec2.internal:8020/user/hadoop/.sparkStaging/application_171416167710
1_0003/hudi-defaults.conf
24/04/26 20:56:35 INFO Client: Uploading resource file:/usr/lib/spark/python/lib/pyspark.zip -> hdfs://ip-172-31-69-133.ec2.internal:8020/user/hadoop/.sparkStaging/application_171
4161677101_0003/pyspark.zip
24/04/26 20:56:35 INFO Client: Uploading resource file:/usr/lib/spark/python/lib/py4j-0.10.9.7-src.zip -> hdfs://ip-172-31-69-133.ec2.internal:8020/user/hadoop/.sparkStaging/application_171
4161677101_0003/py4j-0.10.9.7-src.zip
24/04/26 20:56:36 INFO Client: Uploading resource file:/mnt/tmp/spark-4db2f6a-0ddf-4955-bde0-c52a65fe4c83/_spark_conf_1049270680638660577.zip -> hdfs://ip-172-31-69-133.ec2.internal:8020
24/04/26 20:56:36 INFO Client: Uploading resource file:/etc/spark/conf.dist/_spark_conf_.zip -> hdfs://ip-172-31-69-133.ec2.internal:8020/user/hadoop/.sparkStaging/application_171
4161677101_0003/_spark_conf_.zip
24/04/26 20:56:36 INFO SecurityManager: Changing view acls to: hadoop
24/04/26 20:56:36 INFO SecurityManager: Changing modify acls to: hadoop
24/04/26 20:56:36 INFO SecurityManager: Changing view acls groups to:
24/04/26 20:56:36 INFO SecurityManager: Changing modify acls groups to:
24/04/26 20:56:36 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: hadoop; groups with view permissions: EMPTY; users with modi
ty permissions: hadoop; groups with modify permission: EMPTY
24/04/26 20:56:36 INFO Client: Submitting application application_1714161677101_0003 to ResourceManager
24/04/26 20:56:36 INFO ApplicationClientImpl: Submitted application application_1714161677101_0003
24/04/26 20:56:37 INFO Client: Application report for application_1714161677101_0003 (state: ACCEPTED)
24/04/26 20:56:37 INFO Client:
  client token: N/A
  diagnostics: AM container is launched, waiting for AM container to Register with RM
  ApplicationMaster host: N/A
  ApplicationMaster RPC port: -1
  queue: default
  start time: 1714164996129
  final status: UNDEFINED
  tracking URL: http://ip-172-31-69-133.ec2.internal:20888/proxy/application_1714161677101_0003/
  user: hadoop
24/04/26 20:56:40 INFO YarnClientSchedulerBackend: Add WebUI Filter. org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter, Map(PROXY_HOSTS -> ip-172-31-69-133.ec2.internal, PROXY_URI_
BASE -> http://ip-172-31-69-133.ec2.internal:20888/proxy/application_1714161677101_0003, /proxy/application_1714161677101_0003)
24/04/26 20:56:41 INFO Client: Application report for application_1714161677101_0003 (state: RUNNING)
24/04/26 20:56:41 INFO Client:
  client token: N/A
  diagnostics: N/A
  ApplicationMaster host: 172.31.79.179
  ApplicationMaster RPC port: -1
  queue: default
  start time: 1714164996129
  final status: UNDEFINED
  tracking URL: http://ip-172-31-69-133.ec2.internal:20888/proxy/application_1714161677101_0003/
```

```

user: hadoop
24/04/26 20:56:41 INFO YarnClientSchedulerBackend: Application application_1714161677101_0003 has started running.
24/04/26 20:56:41 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 40265.
24/04/26 20:56:41 INFO NettyBlockTransferService: Server created on ip-172-31-69-133.ec2.internal:40265
24/04/26 20:56:41 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
24/04/26 20:56:41 INFO BlockManager: external shuffle service port = 7337
24/04/26 20:56:41 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, ip-172-31-69-133.ec2.internal, 40265, None)
24/04/26 20:56:41 INFO BlockManager: Registered BlockManager BlockManagerId(driver, ip-172-31-69-133.ec2.internal, 40265, None)
24/04/26 20:56:41 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, ip-172-31-69-133.ec2.internal, 40265, None)
24/04/26 20:56:41 INFO SingleEventLogFileWriter: Logging events to hdfs://was/log/spark/apps/application_1714161677101_0003.inprogress
24/04/26 20:56:41 INFO Utils: Using 50 preallocated executors [minExecutors: 0]. Set spark.dynamicAllocation.preallocateExecutors to 'false' disable executor preallocation.
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /jobs: org.apache.hadoop.yarn.server.webproxy.amfilter.AmfIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /jobs/job: org.apache.hadoop.yarn.server.webproxy.amfilter.AmfIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /jobs/job/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmfIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /stages: org.apache.hadoop.yarn.server.webproxy.amfilter.AmfIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /stages/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmfIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /stages/stage: org.apache.hadoop.yarn.server.webproxy.amfilter.AmfIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /stages/stage/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmfIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /stagepool: org.apache.hadoop.yarn.server.webproxy.amfilter.AmfIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /stagepool/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmfIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /stages/kill: org.apache.hadoop.yarn.server.webproxy.amfilter.AmfIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /storage/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmfIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /storage/rdd: org.apache.hadoop.yarn.server.webproxy.amfilter.AmfIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /storage/rdd/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmfIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /environment: org.apache.hadoop.yarn.server.webproxy.amfilter.AmfIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /environment/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmfIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /executors: org.apache.hadoop.yarn.server.webproxy.amfilter.AmfIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /executors/threadDump: org.apache.hadoop.yarn.server.webproxy.amfilter.AmfIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /executors/threadDump/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmfIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /executors/heapHistogram: org.apache.hadoop.yarn.server.webproxy.amfilter.AmfIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /executors/heapHistogram/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmfIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /static: org.apache.hadoop.yarn.server.webproxy.amfilter.AmfIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /: org.apache.hadoop.yarn.server.webproxy.amfilter.AmfIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /api: org.apache.hadoop.yarn.server.webproxy.amfilter.AmfIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /metrics: org.apache.hadoop.yarn.server.webproxy.amfilter.AmfIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /metrics/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmfIpFilter
24/04/26 20:56:41 INFO YarnClientSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.0
Importing: s3a://dataset-programming-assignment-2/TrainingDataset.csv
>>> Model Path set: 3a7/dataset-programming-assignment-2/models
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
root
|--- ****"fixed acidity": string (nullable = true)
|--- ****"volatile acidity": string (nullable = true)
|--- ****"citric acid": string (nullable = true)
|--- ****"residual sugar": string (nullable = true)
|--- ****"chlorides": string (nullable = true)
|--- ****"free sulfur dioxide": string (nullable = true)
|--- ****"total sulfur dioxide": string (nullable = true)

```

```

SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
root
|--- ****"fixed acidity": string (nullable = true)
|--- ****"volatile acidity": string (nullable = true)
|--- ****"citric acid": string (nullable = true)
|--- ****"residual sugar": string (nullable = true)
|--- ****"chlorides": string (nullable = true)
|--- ****"free sulfur dioxide": string (nullable = true)
|--- ****"total sulfur dioxide": string (nullable = true)
|--- ****"density": string (nullable = true)
|--- ****"pH": string (nullable = true)
|--- ****"sulphates": string (nullable = true)
|--- ****"alcohol": string (nullable = true)
|--- ****"quality": string (nullable = true)
+-----+-----+-----+-----+-----+-----+-----+
|---****"fixed acidity"|"---"volatile acidity"|"---"citric acid"|"---"residual sugar"|"---"chlorides"|"---"free sulfur dioxide"|"---"total sulfur dioxide"|"---"density"|"---"pH
|---****"sulphates"|"---"alcohol"|"---"quality"
+-----+-----+-----+-----+-----+-----+-----+
| 8.5| 0.22| 0.48| 1.0| 0.077| 29| 60| 0.9968|
3.39| 0.53| 9.4| 6| 0.31| 2.3| 0.082| 23| 71| 0.9982|
| 7.6| 0.65| 9.7| 5| 0.21| 1.6| 0.106| 10| 37| 0.9966|
3.52| 7.9| 0.43| 5| 0.11| 2.3| 0.084| 9| 67| 0.9968|
3.17| 0.91| 9.5| 5| 0.14| 2.4| 0.085| 21| 40| 0.9968|
3.17| 0.53| 9.4| 5| 0.41| 0.14| 0.085| 21| 40| 0.9968|
| 6.9| 0.63| 9.7| 6| 0.39| 0.16| 0.081| 11| 23| 0.9955|
3.43| 6.3| 0.56| 9.3| 0.41| 5| 0.24| 1.4| 0.081| 4| 11| 0.9962|
| 7.6| 0.59| 9.5| 5| 0.43| 5| 0.21| 1.8| 0.081| 4| 11| 0.9962|
3.28| 7.9| 0.59| 9.5| 0.43| 5| 0.21| 1.6| 0.106| 10| 37| 0.9966|
3.17| 0.59| 9.5| 5| 0.71| 0| 0| 1.9| 0.081| 14| 35| 0.9972|
3.47| 7.1| 0.55| 9.4| 0.645| 5| 0| 2| 0.082| 8| 16| 0.9964|
| 7.8| 0.59| 9.8| 6| 0| 0| 2| 0.082| 8| 16| 0.9958|
3.38| 6.7| 0.54| 10.1| 0.675| 5| 0.07| 2.4| 0.089| 17| 82| 0.9958|
| 6.9| 0.54| 10.1| 5| 0.685| 0| 0| 2.5| 0.105| 22| 37| 0.9966|
3.46| 0.57| 10.6| 6| 0.655| 0| 0.12| 2.3| 0.083| 15| 113| 0.9966|
| 8.3| 0.66| 9.8| 5| 0.655| 0| 0.12| 2.3| 0.083| 15| 113| 0.9966|
3.17| 6.9| 0.52| 9.4| 0.605| 6| 0.12| 10.7| 0.073| 40| 83| 0.9993|
3.45| 5.2| 0.55| 9.2| 0.32| 0.25| 1.8| 0.103| 13| 50| 0.9957|
3.38| 5.2| 0.55| 9.2| 0.32| 0.25| 1.8| 0.103| 13| 50| 0.9957|
| 7.8| 0.55| 9.6| 6| 0.645| 0| 0| 5.5| 0.086| 5| 18| 0.9986|
3.4| 7.8| 0.55| 9.6| 0.6| 0.14| 2.4| 0.086| 3| 15| 0.9975|

```

The results will be displayed here, along with the Accuracy and F1 scores of the Machine Learning methods used.

```

Training Data Set Metrics
Accuracy: 0.6114151681000782
F-measure: 0.5904050519731796
+-----+-----+-----+-----+-----+-----+-----+-----+
|fixed acidity|volatile acidity|citric acid|residual sugar|chlorides|free sulfur dioxide|total sulfur dioxide|density| pH|sulphates|alcohol|label|
+-----+-----+-----+-----+-----+-----+-----+-----+
|    7.4|      0.7|     0.0|     1.9|   0.076|     1.1|     0.9978| 13.51|   0.56|     9.4|   5.0|
|    7.8|      0.88|    0.0|     2.6|   0.098|     2.5|     0.9968| 13.2|   0.68|     9.8|   5.0|
|    7.8|      0.76|    0.04|     2.3|   0.092|     1.5|     0.997| 13.26|   0.65|     9.8|   5.0|
|   11.2|      0.28|    0.56|     1.9|   0.075|     1.7|     0.998| 13.16|   0.58|     9.8|   6.0|
|    7.4|      0.7|     0.0|     1.9|   0.076|     1.1|     0.9978| 13.51|   0.56|     9.4|   5.0|
+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 5 rows

Validation Training Set Metrics
+-----+-----+-----+
|features|label|prediction|
+-----+-----+-----+
|[9.4, 0.56, 3.51, 0.9978, 11.0, 34.0, 0.0, 0.076, 1.9, 0.0, 0.0, 0.7, 7.4]| 15.0 | 15.0 |
|[9.8, 0.68, 3.2, 0.9968, 25.0, 67.0, 0.0, 0.098, 2.6, 0.0, 0.0, 0.88, 7.8]| 15.0 | 15.0 |
|[9.8, 0.65, 3.26, 0.997, 15.0, 54.0, 0.0, 0.092, 2.3, 0.04, 0.76, 7.8]| 15.0 | 15.0 |
|[9.8, 0.58, 3.16, 0.998, 17.0, 60.0, 0.0, 0.075, 1.9, 0.0, 0.56, 0.28, 11.2]| 16.0 | 15.0 |
|[9.4, 0.56, 3.51, 0.9978, 11.0, 34.0, 0.0, 0.076, 1.9, 0.0, 0.0, 0.7, 7.4]| 15.0 | 15.0 |
+-----+-----+-----+
only showing top 5 rows

The accuracy of the model is 0.575
F1: 0.5619407071339173

```

DOCKER IMPLEMENTATION–

1.Update System Packages:

sudo yum update -y 2.Install

Docker:

sudo yum install -y docker

3.Start Docker Service:

sudo service docker start 4.Check

Docker Service Status: sudo

service docker status

```

Dependencies resolved.
Nothing to do.
Complete!
[ec2-user@ip-172-31-47-91 ~]$ sudo service docker start
Redirecting to /bin/systemctl start docker.service
[ec2-user@ip-172-31-47-91 ~]$ sudo service docker status
Redirecting to /bin/systemctl status docker.service
● docker.service - Docker Application Container Engine
   Loaded: loaded (/usr/lib/systemd/system/docker.service; disabled; preset: disabled)
     Active: active (running) since Sun 2024-04-28 18:40:47 UTC; 1min 4s ago
TriggeredBy: ● docker.socket
   Docs: https://docs.docker.com
 Process: 25902 ExecStartPre=/bin/mkdir -p /run/docker (code=exited, status=0/SUCCESS)
 Process: 25903 ExecStartPre=/usr/libexec/docker/docker-setup-runtimes.sh (code=exited, status=0/SUCCESS)
 Main PID: 25904 (dockerd)
   Tasks: 10
     Memory: 107.8M
        CPU: 331ms
      CGroup: /system.slice/docker.service
              └─25904 /usr/bin/dockerd -H fd:// --containerd=/run/containerd/containerd.sock --default-ulimit nofile=32768:65536

Apr 28 18:40:46 ip-172-31-47-91.us-east-2.compute.internal systemd[1]: Starting docker.service - Docker Application Container Engine...
Apr 28 18:40:46 ip-172-31-47-91.us-east-2.compute.internal dockerd[25904]: time="2024-04-28T18:40:46.670614846Z" level=info msg="Starting up"
Apr 28 18:40:46 ip-172-31-47-91.us-east-2.compute.internal dockerd[25904]: time="2024-04-28T18:40:46.605514723Z" level=info msg="Loading containers: start."
Apr 28 18:40:47 ip-172-31-47-91.us-east-2.compute.internal dockerd[25904]: time="2024-04-28T18:40:47.206510575Z" level=info msg="Loading containers: done."
Apr 28 18:40:47 ip-172-31-47-91.us-east-2.compute.internal dockerd[25904]: time="2024-04-28T18:40:47.247476769Z" level=info msg="Docker daemon" commit=f417435 containerd-snapshotter=false
Apr 28 18:40:47 ip-172-31-47-91.us-east-2.compute.internal dockerd[25904]: time="2024-04-28T18:40:47.247666950Z" level=info msg="Daemon has completed initialization"
Apr 28 18:40:47 ip-172-31-47-91.us-east-2.compute.internal dockerd[25904]: time="2024-04-28T18:40:47.286678790Z" level=info msg="API listen on /run/docker.sock"
Apr 28 18:40:47 ip-172-31-47-91.us-east-2.compute.internal systemd[1]: Started docker.service - Docker Application Container Engine.

```

5. Create a Dockerfile and build an image with the docker build command.

```
sudo docker build -t vd348/cs643-programming-assignment-2 .
```

6. To check if a Docker image was built, use the following command:

```
# sudo docker image ls
```

```
[root@ip-172-31-68-235 CS643_Programming_assignment_2]# docker image ls
REPOSITORY          TAG      IMAGE ID   CREATED             SIZE
ga348/cs643-programming-assignment-2  latest   7fc54dac10b9  19 seconds ago  2.42GB
[root@ip-172-31-68-235 CS643_Programming_assignment_2]#
```

You can see here that your docker image has been built.

7. To run the docker image, use the following command:

```
sudo docker run -it vd348/cs643-programming-assignment-2
```

Instead of using the image name, you can use the image ID: # sudo docker run -it

8. This results in the same return for Accuracy and F1 scores.

Wine Quality Prediction Data										
Root Node		Level 1			Level 2			Level 3		
		fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	ph
id	quality	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	ph
1	1	8.9	0.22	0.48	1.8	0.077	29	60	0.9968	3.91
2	1	0.53	9.4	0.39	6	0.31	2.3	0.082	23	71
3	1	7.6	0.65	9.7	5	0.21	1.6	0.106	10	37
4	1	0.65	7.9	0.43	5	0.11	2.3	0.084	9	67
5	1	0.91	9.5	0.49	5	0.14	2.4	0.085	21	40
6	1	8.5	0.53	9.4	5	0.41	1.4	0.081	11	23
7	1	6.9	0.63	9.7	6	0.16	0.8	0.081	4	11
8	1	6.3	0.56	9.3	5	0.24	1.6	0.106	10	37
9	1	7.6	0.59	9.5	5	0.21	1.6	0.106	10	37
10	1	7.9	0.91	9.5	5	0.71	0	1.9	0.081	14
11	1	7.1	0.55	9.4	5	0	2.1	0.082	8	16
12	1	7.8	0.59	9.8	6	0.07	2.4	0.089	17	82
13	1	6.7	0.54	10.1	5	0	2.5	0.105	22	37
14	1	6.9	0.54	10.1	5	0.685	0	1.05	15	113
15	1	8.3	0.57	10.6	6	0.12	2.3	0.083	113	0.9966
16	1	6.9	0.66	9.8	5	0.12	10.7	0.073	40	83
17	1	6.9	0.52	9.4	6	0.605	0	0.073	13	50
18	1	5.2	0.55	9.2	5	0.32	0.25	1.8	0.103	18
19	1	7.8	0.55	9.2	5	0	5.5	0.086	5	0.9957
20	1	7.8	0.55	9.6	6	0.6	0.14	2.4	0.086	3
21	1	7.8	0.55	9.6	6	0.14	2.4	0.086	3	15
22	1	7.8	0.55	9.6	6	0.14	2.4	0.086	3	0.9975

9. Run the following command to submit the produced Docker image to the DockerHub repository:

```
sudo docker push vd348zcs643-programming-assignment-2
```

```
Log in with your Docker ID or email address to push and pull images from Docker Hub. If you don't have a Docker ID, head over to https://hub.docker.com/ to create one. You can log in with your password or a Personal Access Token (PAT). Using a limited-scope PAT grants better security and is required for organizations using SSO. Learn more at docker.com/go/access-tokens

Username: venkanna369
Password:
WARNING! Your password will be stored unencrypted in /root/.docker/config.json. Configure a credential helper to remove this warning. See https://docs.docker.com/engine/reference/commandline/login/#credentials-store

Login Succeeded
[root@ip-172-31-47-91 ~]# sudo docker push venkanna369/wine_production
Using default tag: latest
The push refers to repository [docker.io/venkanna369/wine_production]
242b6ead8f3e: Pushed
4ac4c1912604: Pushed
d569805bd5bc: Pushed
b154a2967a0b: Pushed
211fd1185450: Pushed
57c651240c9f: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
553c43e2f0d1: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
36ef902c4c66: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
e1a1b88cc1ff8: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
632ccc24d10f: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
0933d669b084: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
367158596a5c: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
c9ac6abb0c04d: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
41caa71c39b5: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
97393f8c8163: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
d7802b8508aaf: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
e3abdc2e9252: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
earfe6e032dbd: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
92a4e8a3140f: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
latest: digest: sha256:2c2704c4809b18c3b88d762af58aclfd86fb5f9da9a317f29a5bb93816794fc8 size: 4516
```

10. Download and execute the Docker image from the DockerHub repository, following the instructions provided on the website.

Git Bash:

the steps for moving code from my local system to a GitHub repository using Git Bash.

1. git init
2. git add.

3. git status

4. git commit -m " updated code"

5. git remote add origin https://github.com/gayatriaavula/winepred_quality.git

6. git push -u origin main

```
aavu1_nutuq9v@AAVULADURGA MINGW64 ~/OneDrive/Documents/DevOps/cloud
$ git init
Initialized empty Git repository in C:/Users/aavu1_nutuq9v/OneDrive/documents/DevOps/cloud/.git/
aavu1_nutuq9v@AAVULADURGA MINGW64 ~/OneDrive/Documents/DevOps/cloud (master)
$ git status
On branch master

No commits yet

Untracked files:
  (use "git add <file>..." to include in what will be committed)
    1.png
    2.png
    3.png
    4.png
    CS643_homeworkset4#_Gayatri_Aavula.docx
    Internship offer letter.docx
    aws-keys.txt
    aws-programming1-notes.txt
    certificate_226002383_9fe5093e.pdf
    images.docx
    midi.txt
    object.ppm
    object1.ppk
    presentation_cloud.docx
    programming2/
    recording-images.docx
    winepred_main.py/

```

```
aavu1_nutuq9v@AAVULADURGA MINGW64 ~/OneDrive/Documents/DevOps/cloud/winepred_quality (main)
$ git status
On branch main
Your branch is up to date with 'origin/main'.

Untracked files:
  (use "git add <file>..." to include in what will be committed)
    Dockerfile
    TrainingDataset.csv
    ValidationDataset.csv
    WineTesting.py
    WineTraining.py

nothing added to commit but untracked files present (use "git add" to track)

aavu1_nutuq9v@AAVULADURGA MINGW64 ~/OneDrive/Documents/DevOps/cloud/winepred_quality (main)
$ git add .
aavu1_nutuq9v@AAVULADURGA MINGW64 ~/OneDrive/Documents/DevOps/cloud/winepred_quality (main)
$ git status
On branch main
Your branch is up to date with 'origin/main'.

Changes to be committed:
  (use "git restore --staged <file>..." to unstaged)
    new file:   Dockerfile
    new file:   TrainingDataset.csv
    new file:   ValidationDataset.csv
    new file:   WineTesting.py
    new file:   WineTraining.py

aavu1_nutuq9v@AAVULADURGA MINGW64 ~/OneDrive/Documents/DevOps/cloud/winepred_quality (main)
$ git git commit -m "updated code"
git: 'git' is not a git command. See 'git --help'.
The most similar command is
  init

aavu1_nutuq9v@AAVULADURGA MINGW64 ~/OneDrive/Documents/DevOps/cloud/winepred_quality (main)
$ git commit -m "updated code"
[main c041357] updated code
 5 files changed, 1682 insertions(+)
 create mode 100644 Dockerfile
 create mode 100644 TrainingDataset.csv
 create mode 100644 ValidationDataset.csv
 create mode 100644 WineTesting.py
 create mode 100644 WineTraining.py

aavu1_nutuq9v@AAVULADURGA MINGW64 ~/OneDrive/Documents/DevOps/cloud/winepred_quality (main)
$ l
```

```
aavu1_nutuq9v@AAVULADURGA MINGW64 ~/OneDrive/Documents/DevOps/cloud/winepred_quality (main)
$ git remote add origin https://github.com/gayatriaavula/winepred_quality.git
error: remote origin already exists.

aavu1_nutuq9v@AAVULADURGA MINGW64 ~/OneDrive/Documents/DevOps/cloud/winepred_quality (main)
$ git push -u origin main
Enumerating objects: 8, done.
Counting objects: 100% (8/8), done.
Delta compression using up to 12 threads
Compressing objects: 100% (7/7), done.
Writing objects: 100% (7/7), 24.54 KiB | 2.73 MiB/s, done.
Total 7 (delta 0), reused 0 (delta 0), pack-reused 0
To https://github.com/gayatriaavula/winepred_quality.git
  f8dd3c3..c041357 main -> main
branch 'main' set up to track 'origin/main'.

aavu1_nutuq9v@AAVULADURGA MINGW64 ~/OneDrive/Documents/DevOps/cloud/winepred_quality (main)
$ |
```

