

CS643 Cloud Programming Assignment 2

READme file:

Git Hub link:

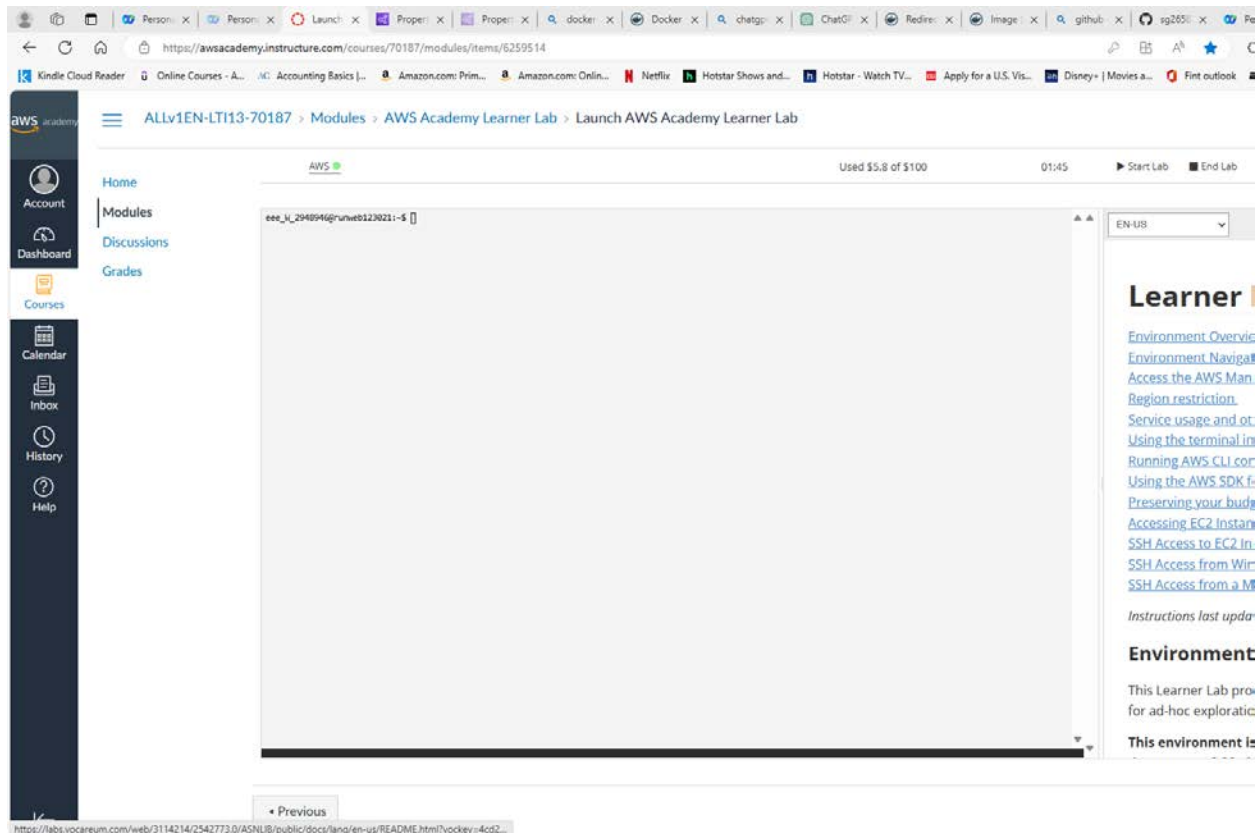
https://github.com/sg2658/cloud_programming_assignment_2

DOCKER HUB link:

sg2658/wine_quality_prediction_tags | Docker Hub

Steps for the Execution:

1. Create a Key-pair for the EMR Cluster
2. Create an S3 bucket
Created an S3 bucket in aws: emrbucketone
3. Then go to EMR console and create EMR cluster: EMRCluster
4. Creating the spark in the AWS instance by using EMR console



EMR cluster:

Amazon EMR

EMR Serverless

EMR on EC2

- Clusters
- Notebooks and Git repos
- Events
- Block public access
- Security configurations

EMR on EKS

- Virtual clusters

EMR Studio

- Getting Started
- Studios
- Workspaces (Notebooks)

What's New

- Video tour

Compact mode

Amazon EMR > EMR on EC2: Clusters

Clusters (4) Info

Filter clusters by status

Find clusters

Filter clusters by creation date-time

	Cluster ID	Cluster name	Status	Creation time (UTC-04:00)	Elapsed time	Ne
<input type="checkbox"/>	j-2WS3PPSSF3M7L	Wine_Quality	Waiting Ready to run steps	April 28, 2024, 20:27	2 hours	4E
<input type="checkbox"/>	j-DLWHKZUX053Z	Wine_Quality	Terminated with errors Instance failure	April 28, 2024, 20:20	1 minute	0
<input type="checkbox"/>	j-1TVL66MPIQM2X	Wine_Quality	Terminated User request	April 28, 2024, 20:17	3 minutes, 14 seconds	0
<input type="checkbox"/>	j-3ESRQ1XELL90U	My cluster	Terminated User request	April 26, 2024, 01:48	1 hour, 9 minutes	9E

CloudShell

Feedback

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Kindle Cloud Reader Online Courses - Accounting Basics Amazon.com: Prim... Amazon.com: Onlin... Netflix Hotstar Shows and... Hotstar - Watch TV... Apply for a U.S. Vis... Disney+ | Movies a... Fint outlook

Services Search [Alt+S]

EC2 S3 Simple Queue Service

Amazon EMR > EMR on EC2: Clusters > Wine_Quality

Wine_Quality Updated less than a minute ago Terminate

▼ Summary

Cluster info	Applications	Cluster management	Status and time
Cluster ID j-2W53PP5SF3M7L	Amazon EMR version emr-7.1.0	Log destination in Amazon S3 aws-logs-975050278546-us-east-1/elasticmapreduce	Status Waiting
Cluster configuration Instance groups	Installed applications Hadoop 3.3.6, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.8.0, Spark 3.5.0	Persistent application UIs Spark History Server YARN timeline server Tez UI	Creation time April 28, 2024, 20:27 (UTC)
Capacity 1 Primary 1 Core 4 Task		Primary node public DNS ec2-3-231-158-105.compute-1.amazonaws.com Connect to the Primary node using SSH Connect to the Primary node using SSM	Elapsed time 2 hours, 1 minute

Properties Bootstrap actions Instances (Hardware) Steps Applications Configurations Monitoring Events Tags (0)

Operating system Info	Cluster logs Info	Cluster termination and node repla
Amazon Linux release 2023.4.20240416.0	Archive log files to Amazon S3 Turned on Amazon S3 location s3://aws-logs-975050278546-us-east-1/elasticmapreduce/	Termination option Manually terminate cluster Termination protection Off

Network and security Info

CloudShell Feedback © 2024, Amazon Web Services, Inc. or its a

Bucket creation:

Bucket name: emrbucketone

Amazon S3 console screenshot showing the 'emrbucketone' bucket. The left sidebar lists navigation options: Buckets, Access Grants, Access Points, Object Lambda Access Points, Multi-Region Access Points, Batch Operations, IAM Access Analyzer for S3, Storage Lens, Dashboards, Storage Lens groups, AWS Organizations settings, Feature spotlight, and AWS Marketplace for S3. The main content area displays the bucket name 'emrbucketone' and tabs for Objects, Properties, Permissions, Metrics, Management, and Access Points. The 'Objects' tab is active, showing a list of 5 objects:

<input type="checkbox"/>	Name	Type	Last modified	Size	5s
<input type="checkbox"/>	Dockerfile	-	April 28, 2024, 20:25:56 (UTC-04:00)	1.8 KB	5s
<input type="checkbox"/>	prediction.py	py	April 28, 2024, 20:25:56 (UTC-04:00)	2.4 KB	5s
<input type="checkbox"/>	training.py	py	April 28, 2024, 20:25:57 (UTC-04:00)	4.0 KB	5s
<input type="checkbox"/>	TrainingDataset.csv	csv	April 28, 2024, 20:25:58 (UTC-04:00)	67.2 KB	5s
<input type="checkbox"/>	ValidationDataset.csv	csv	April 28, 2024, 20:25:57 (UTC-04:00)	8.6 KB	5s

At the bottom of the console, there is a footer with 'CloudShell' and 'Feedback' links, and a copyright notice: '© 2024, Amazon Web Services, Inc. or its affiliates'.

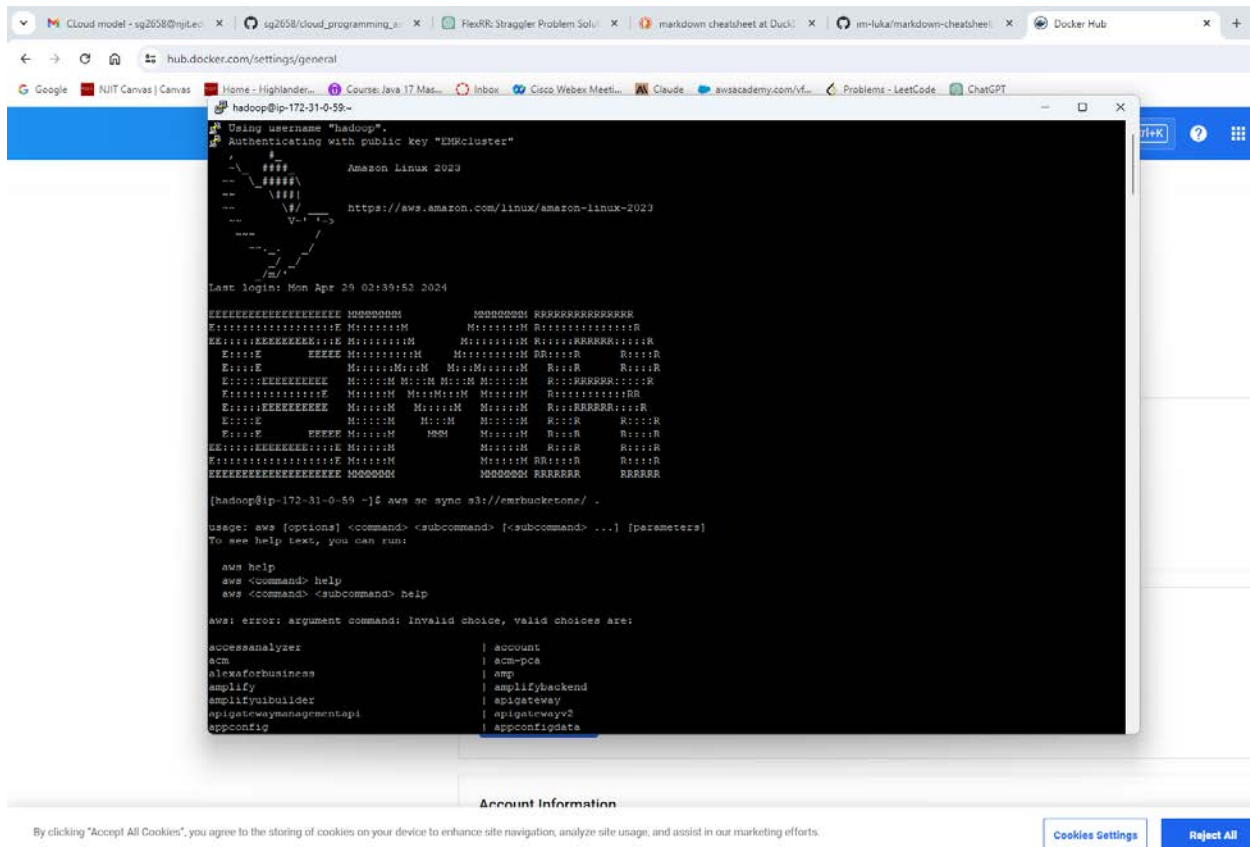
EC2 instance-master:

The screenshot shows the AWS Management Console interface for the EC2 service. The left sidebar contains navigation links for various AWS services, including EC2, S3, and IAM. The main content area displays the 'Instances (6)' page, which shows a list of EC2 instances. The instances are filtered by 'Instance state = running'. The table below shows the details of the instances, including their public IP addresses, monitoring status, and security groups.

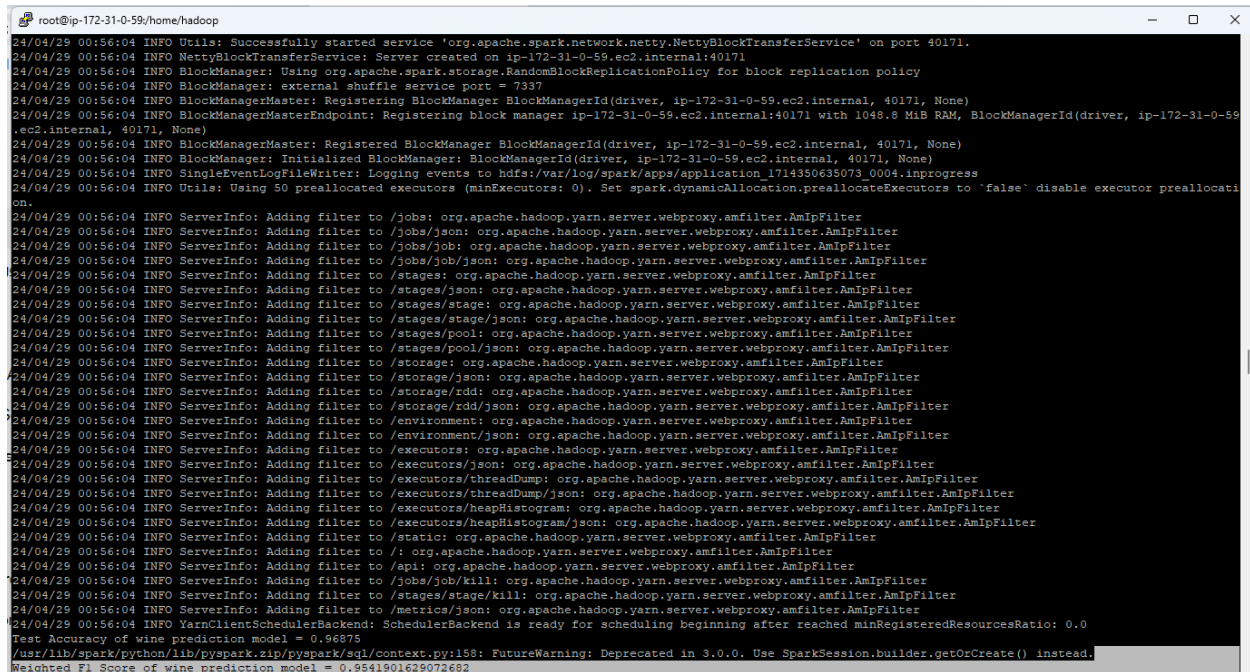
Zone	Public IPv4 DNS	Public IPv4	Elastic IP	IPv6 IPs	Monitoring	Security group name	Key name	Launch time
ec2-44-213-104-56.co...	44.213.104.56	-	-	disabled	ElasticMapReduce-slave	EMRcluster	2024/04/28	
ec2-3-231-158-105.co...	3.231.158.105	-	-	disabled	ElasticMapReduce-master	EMRcluster	2024/04/28	
ec2-3-239-172-62.co...	3.239.172.62	-	-	disabled	ElasticMapReduce-slave	EMRcluster	2024/04/28	
ec2-44-199-199-30.co...	44.199.199.30	-	-	disabled	ElasticMapReduce-slave	EMRcluster	2024/04/28	
ec2-44-222-221-197.co...	44.222.221.197	-	-	disabled	ElasticMapReduce-slave	EMRcluster	2024/04/28	
ec2-3-238-148-114.co...	3.238.148.114	-	-	disabled	ElasticMapReduce-slave	EMRcluster	2024/04/28	

We are training an ML model on an EC2 instance in a spark cluster without Docker:

1.Now the cluster will accept the tasks to run the ML model



WITHOUT Docker:



Now we are running ML model using the Docker:

1. Create a docker account and sign up.
2. After the successful login, download and setup the docker in your local system
3. Install the docker
4. Login the docker in the power shell by the command

docker login

username:

Pwd:

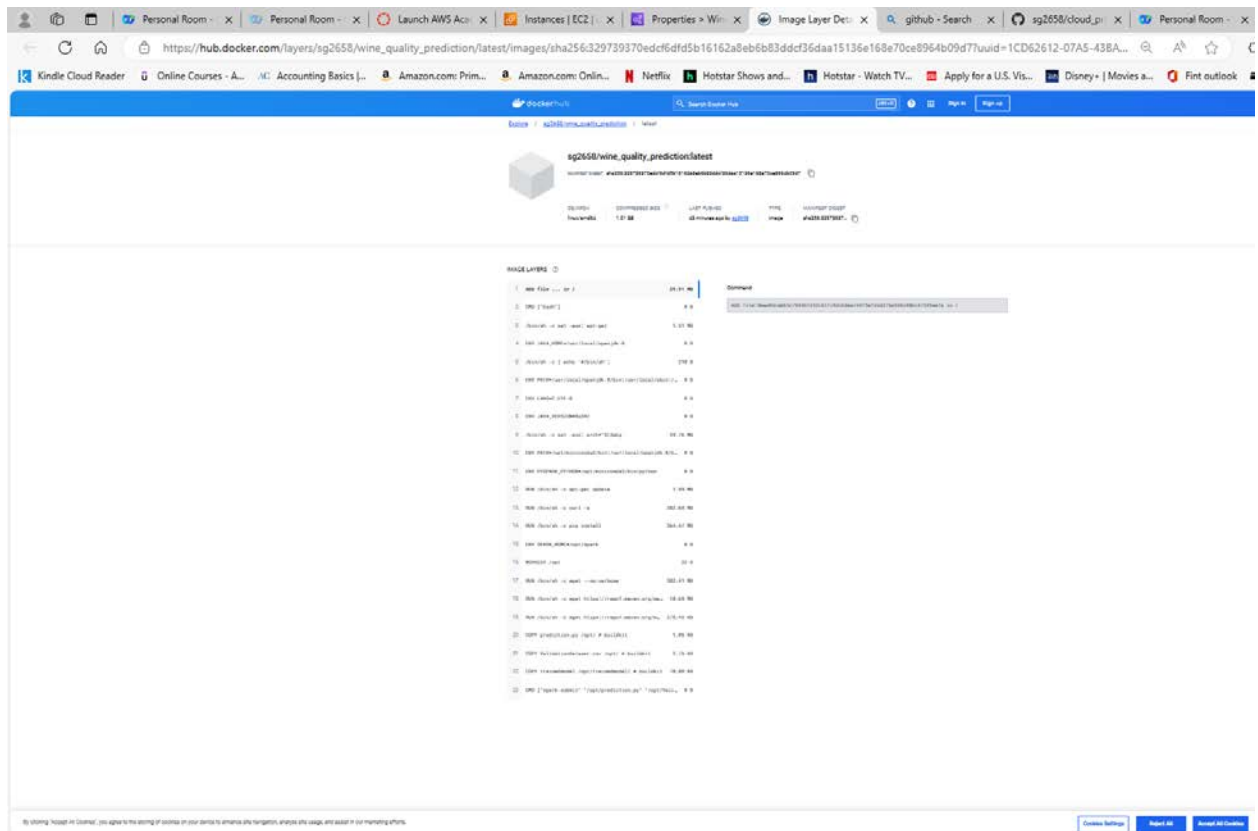
With DOCKER:

```
24/04/29 01:17:20 INFO Utils: Successfully started service 'SparkUI' on port 4040.
24/04/29 01:17:20 INFO Executor: Starting executor ID driver on host 2605a4a95061
24/04/29 01:17:20 INFO Executor: OS info Linux, 6.1.84-99.169.amzn2023.x86_64, amd64
24/04/29 01:17:20 INFO Executor: Java version 1.8.0_342
24/04/29 01:17:20 INFO Executor: Starting executor with user classpath (userClassPathFirst = false): ''
24/04/29 01:17:20 INFO Executor: Created or updated repl class loader org.apache.spark.util.MutableURLClassLoader@2153aa7c for default.
24/04/29 01:17:20 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 41057.
24/04/29 01:17:20 INFO NettyBlockTransferService: Server created on 2605a4a95061:41057
24/04/29 01:17:20 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
24/04/29 01:17:20 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, 2605a4a95061, 41057, None)
24/04/29 01:17:20 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, 2605a4a95061, 41057, None)
24/04/29 01:17:20 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, 2605a4a95061, 41057, None)
Test Accuracy of wine prediction model = 0.96875
/opt/spark/python/lib/pyspark.zip/pyspark/sql/context.py:158: FutureWarning: Deprecated in 3.0.0. Use SparkSession.builder.getOrCreate() instead.
Weighted F1 Score of wine prediction model = 0.9541901629072682
Exiting Spark Application
```

Docker hub:

The screenshot shows the Docker Hub interface for the image `sg2658/wine_quality_prediction:latest`. The page includes a search bar, navigation links, and a detailed view of the image. The image details section shows the manifest digest, OS/ARCH (linux/amd64), compressed size (1.01 GB), last pushed time (43 minutes ago), and type (Image). Below this, the 'IMAGE LAYERS' section lists the layers of the image, including the base image 'ADD file ... in /' and various system and application files. The command to run the image is also displayed.

Layer	Size	Command
1	29.91 MB	ADD file:0eae8dca665c7044bf242cb1fc92cb8ea744f5af2dd376a558c98bc47349aefc in /
2	0 B	CMD ["bash"]
3	1.51 MB	/bin/sh -c set -eux; apt-get
4	0 B	ENV JAVA_HOME=/usr/local/openjdk-8
5	210 B	/bin/sh -c { echo '#/bin/sh';
6	0 B	ENV PATH=/usr/local/openjdk-8/bin:/usr/local/sbin:/
7	0 B	ENV LANG=C.UTF-8
8	0 B	ENV JAVA_VERSION=8u342
9	29.76 MB	/bin/sh -c set -eux; arch=\$(dpkg



GIT hub:

