

CS643 - AWS Spark Wine Quality Prediction Application

Name - Abhiraj Gupta

UCID - ag2672

Section – 861

- I developed a machine learning model to predict wine quality using publicly available data.
- I trained the model in parallel on EC2 instances to improve efficiency.
- I leveraged Docker to create a container image of the trained model, simplifying the deployment process.

Link to github code:

https://github.com/Abhiraj16/CS643_ProgrammingAssignment2

Link to Docker:

<https://hub.docker.com/r/abhiraj1625/wine-eval>

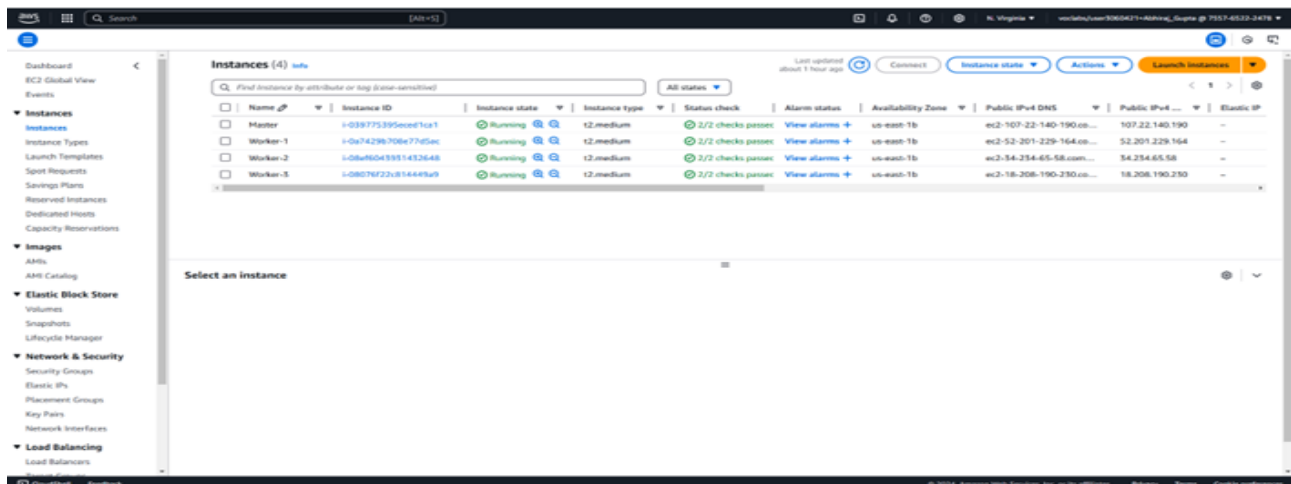
Docker Pull Command:

`docker pull abhiraj1625/wine-eval`

Instructions:

1) Accessing the four instances:

Connect to each of the four instances via SSH using the public IP address and the .pem key file.



2) **Generate SSH Keys on All Instances**

- Create SSH keys on each instance and record the public keys for sharing.

3) **Enable Passwordless SSH**

- Add the public keys from all instances to the `authorized_keys` file on each instance to allow passwordless SSH communication.

4) **Update `/etc/hosts` on All Instances**

- Add the IP addresses and hostnames of all instances to the `/etc/hosts` file to facilitate easier hostname resolution.

5) **Install Java, Maven, and Spark**

- Install Java (OpenJDK 8), Maven, and Spark 3.4.1 on each instance.
- Configure the necessary environment variables to ensure Spark is globally accessible.

6) **Configure Spark Workers**

- Update the workers file in the Spark configuration directory by adding the hostnames or IP addresses of all worker instances.

7) **Set up directories for Training and Evaluation**

- 8) To execute the training code in parallel on all instances, use the `spark-submit` command. This command enables distributed processing by leveraging the Spark cluster setup, allowing the training tasks to run concurrently across all instances.

```
ubuntu@ip-172-31-23-77:~/WinePredTraining/target$ spark-submit --master spark://ip-172-31-23-77.ec2.internal:7077 --class com.example.WineQualityPrediction --deploy-mode cluster wine-quality-prediction-1.0-SNAPSHOT.jar
24/12/09 20:48:30 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
24/12/09 20:48:30 INFO SecurityManager: Changing view acls to: ubuntu
24/12/09 20:48:30 INFO SecurityManager: Changing modify acls to: ubuntu
24/12/09 20:48:30 INFO SecurityManager: Changing view acls groups to:
24/12/09 20:48:30 INFO SecurityManager: Changing modify acls groups to:
24/12/09 20:48:30 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: ubuntu; groups with view permissions: EMPTY; users with modify permissions: ubuntu; groups with modify permissions: EMPTY
24/12/09 20:48:30 INFO Utils: Successfully started service 'driverClient' on port 38213.
24/12/09 20:48:30 INFO TransportClientFactory: Successfully created connection to ip-172-31-23-77.ec2.internal/172.31.23.77:7077 after 45 ms (0 ms spent in bootstraps)
24/12/09 20:48:30 INFO ClientEndpoint: ... waiting before polling master for driver state
24/12/09 20:48:31 INFO ClientEndpoint: Driver successfully submitted as driver-20241209204830-0000
24/12/09 20:48:36 INFO ClientEndpoint: State of driver-20241209204830-0000 is RUNNING
24/12/09 20:48:36 INFO ClientEndpoint: Driver running on 172.31.22.215:34229 (worker-20241209204604-172.31.22.215-34229)
```

Launch AWS Academy Learner

Instance details [EC2] us-east-1

Spark Master at spark://ip-172-...

Not secure107.22.140.190:8080

Spark3.4.1

Spark Master at spark://ip-172-31-23-77.ec2.internal:7077

URL: spark://ip-172-31-23-77.ec2.internal:7077

Alive Workers: 4

Cores in use: 8 Total, 8 Used

Memory in use: 11.3 GiB Total, 5.0 GiB Used

Resources in use:

Applications: 1 Running, 0 Completed

Drivers: 1 Running, 0 Completed

Status: ALIVE

Workers (4)

Worker Id	Address	State	Cores	Memory	Resources
worker-20241209204602-172.31.23.77-37983	172.31.23.77:37983	ALIVE	2 (2 Used)	2.8 GiB (1024.0 MiB Used)	
worker-20241209204604-172.31.22.215-34229	172.31.22.215:34229	ALIVE	2 (2 Used)	2.8 GiB (2.0 GiB Used)	
worker-20241209204605-172.31.28.123-33721	172.31.28.123:33721	ALIVE	2 (2 Used)	2.8 GiB (1024.0 MiB Used)	
worker-20241209204605-172.31.30.180-34089	172.31.30.180:34089	ALIVE	2 (2 Used)	2.8 GiB (1024.0 MiB Used)	

Running Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20241209204835-0000	(kill) Wine Quality Prediction	7	1024.0 MiB		2024/12/09 20:48:35	ubuntu	RUNNING	1 s

Running Drivers (1)

Submission ID	Submitted Time	Worker	State	Cores	Memory	Resources	Main Class	Duration
driver-20241209204830-0000	(kill) 2024/12/09 20:48:30	worker-20241209204604-172.31.22.215-34229	RUNNING	1	1024.0 MiB		com.example.WineQualityPrediction	5 s

Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

Completed Drivers (0)

Submission ID	Submitted Time	Worker	State	Cores	Memory	Resources	Main Class
---------------	----------------	--------	-------	-------	--------	-----------	------------

Launch AWS Academy Learner

Instance details [EC2] us-east-1

Spark Master at spark://ip-172-...

Not secure107.22.140.190:8080

Spark3.4.1

Spark Master at spark://ip-172-31-23-77.ec2.internal:7077

URL: spark://ip-172-31-23-77.ec2.internal:7077

Alive Workers: 4

Cores in use: 8 Total, 0 Used

Memory in use: 11.3 GiB Total, 0.0 B Used

Resources in use:

Applications: 0 Running, 1 Completed

Drivers: 0 Running, 1 Completed

Status: ALIVE

Workers (4)

Worker Id	Address	State	Cores	Memory	Resources
worker-20241209204602-172.31.23.77-37983	172.31.23.77:37983	ALIVE	2 (0 Used)	2.8 GiB (0.0 B Used)	
worker-20241209204604-172.31.22.215-34229	172.31.22.215:34229	ALIVE	2 (0 Used)	2.8 GiB (0.0 B Used)	
worker-20241209204605-172.31.28.123-33721	172.31.28.123:33721	ALIVE	2 (0 Used)	2.8 GiB (0.0 B Used)	
worker-20241209204605-172.31.30.180-34089	172.31.30.180:34089	ALIVE	2 (0 Used)	2.8 GiB (0.0 B Used)	

Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

Running Drivers (0)

Submission ID	Submitted Time	Worker	State	Cores	Memory	Resources	Main Class	Duration
---------------	----------------	--------	-------	-------	--------	-----------	------------	----------

Completed Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20241209204835-0000	Wine Quality Prediction	7	1024.0 MiB		2024/12/09 20:48:35	ubuntu	FINISHED	41 s

Completed Drivers (1)

Submission ID	Submitted Time	Worker	State	Cores	Memory	Resources	Main Class
driver-20241209204830-0000	2024/12/09 20:48:30	worker-20241209204604-172.31.22.215-34229	FINISHED	1	1024.0 MiB		com.example.WineQualityPrediction

Step 9: Build and Push Docker Image

Use the provided Dockerfile to build a Docker image that includes the trained model, validation dataset, and evaluation code. Then, push the Docker image to Docker Hub for

easy access across instances.

Commands:

```
sudo docker build -t abhiraj1625/wine-quality-eval:latest .
```

```
sudo docker push abhiraj1625/wine-quality-eval:latest
```

Step 10: Pull Docker Image on Target Instances

On the desired instances, pull the Docker image from Docker Hub to deploy the evaluation setup.

Command:

```
sudo docker pull abhiraj1625/wine-quality-eval:latest
```

```
ubuntu@ip-172-31-22-77:~/WineEval/target$ docker pull abhiraj1625/wine-eval
Using default tag: latest
latest: Pulling from abhiraj1625/wine-eval
c1d074c93c1c: Pull complete
5e6ec31a09e5: Pull complete
1563ca00434e: Pull complete
f08cbe3372b8: Pull complete
9b2da09d580c: Pull complete
Digest: sha256:54652a11a1b155c6c21dae73661afa7a59bc4527f29dc3ddca621c0a37a2c2
Status: Downloaded newer image for abhiraj1625/wine-eval:latest
docker.io/abhiraj1625/wine-eval:latest
```

Step 11) Start the Docker container on the desired instances to execute the evaluation setup.

```
ubuntu@ip-172-31-22-77:~/WineEval/target$ docker run abhiraj1625/wine-eval
spark 22:05:39.97 INFO ==>
spark 22:05:39.98 INFO ==> Welcome to the Bitnami spark container
spark 22:05:39.98 INFO ==> Subscribe to project updates by watching https://github.com/bitnami/containers
spark 22:05:39.98 INFO ==> Submit issues and feature requests at https://github.com/bitnami/containers/issues
spark 22:05:39.98 INFO ==>

24/12/09 22:05:43 INFO SparkContext: Running Spark version 3.4.1
24/12/09 22:05:43 INFO ResourceUtils: =====
24/12/09 22:05:43 INFO ResourceUtils: No custom resources configured for spark.driver.
24/12/09 22:05:43 INFO ResourceUtils: =====
24/12/09 22:05:43 INFO SparkContext: Submitted application: Wine Quality Evaluation
24/12/09 22:05:43 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 1, script: , vendor: , memory -> name: memory, amount: 1024, script: , vendor: , o
ffHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
24/12/09 22:05:43 INFO ResourceProfile: Limiting resource is cpu
24/12/09 22:05:43 INFO ResourceProfileManager: Added ResourceProfile id: 0
24/12/09 22:05:43 INFO SecurityManager: Changing view acls to: spark
24/12/09 22:05:43 INFO SecurityManager: Changing modify acls to: spark
24/12/09 22:05:43 INFO SecurityManager: Changing view acls groups to:
24/12/09 22:05:43 INFO SecurityManager: Changing modify acls groups to:
24/12/09 22:05:43 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: spark; groups with view permissions: EMPTY; users with modify permissions: spark
; groups with modify permissions: EMPTY
24/12/09 22:05:44 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
24/12/09 22:05:44 INFO Utils: Successfully started service 'sparkDriver' on port 38983.
24/12/09 22:05:44 INFO SparkEnv: Registering MapOutputTracker
24/12/09 22:05:44 INFO SparkEnv: Registering BlockManagerMaster
24/12/09 22:05:44 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
24/12/09 22:05:44 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
24/12/09 22:05:44 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
24/12/09 22:05:44 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-ae5df4e0-2e6f-42ac-bc14-81ea413ced33
24/12/09 22:05:44 INFO MemoryStore: MemoryStore started with capacity 434.4 MiB
```

Finally, I got an F1 Score of 0.8730357142857142, and I have used model as Random Forest.

```
24/12/09 22:05:58 INFO DAGScheduler: Job 8 finished: collectAsMap at MulticlassMetrics.scala:61, took 0.285217 s
F1 Score: 0.8730357142857142
24/12/09 22:05:58 INFO SparkContext: SparkContext is stopping with exitCode 0
24/12/09 22:05:58 INFO SparkUI: Stopped Spark web UI at http://82bbff1a1be9:4040
24/12/09 22:05:58 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
24/12/09 22:05:58 INFO MemoryStore: MemoryStore cleared
24/12/09 22:05:58 INFO BlockManager: BlockManager stopped
24/12/09 22:05:58 INFO BlockManagerMaster: BlockManagerMaster stopped
24/12/09 22:05:58 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
24/12/09 22:05:58 INFO SparkContext: Successfully stopped SparkContext
24/12/09 22:05:58 INFO ShutdownHookManager: Shutdown hook called
24/12/09 22:05:58 INFO ShutdownHookManager: Deleting directory /tmp/spark-561573c7-54d1-43d6-998d-3b98d269d8a8
24/12/09 22:05:58 INFO ShutdownHookManager: Deleting directory /tmp/spark-ba8dc331-c9ef-8a7f-bda7-de5dec140864
```