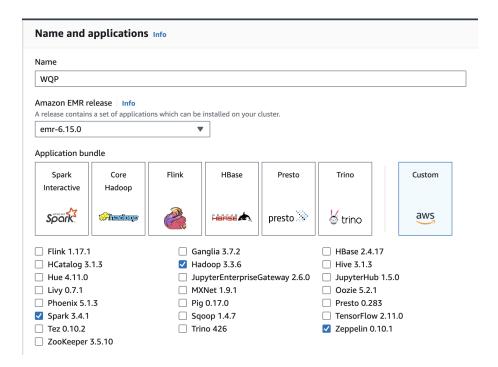
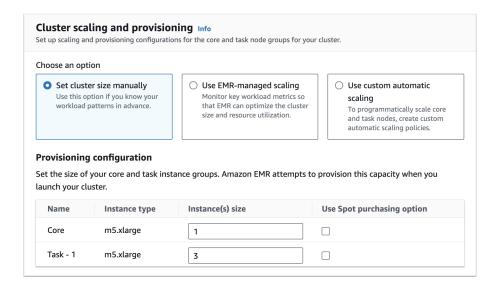
Cluster Creation

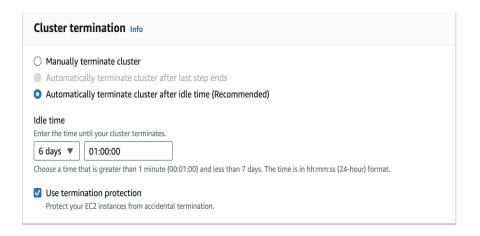
Navigate to Amazon EMR in AWS > Create the Cluster, and select the following apps:



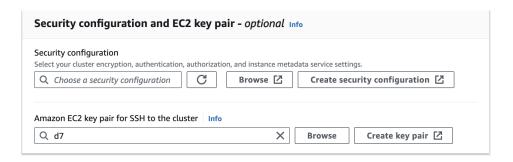
Had to train model using 4 EC2 instances, so configured 'Provisioning Configuration as below:



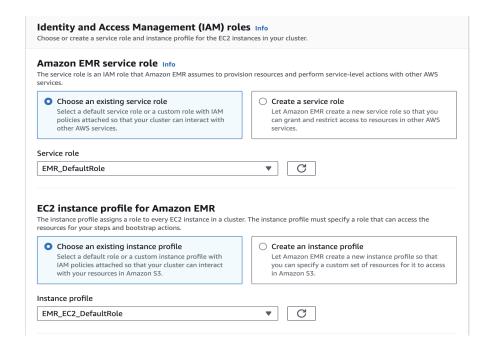
Went with Recommended Cluster Termination settings:



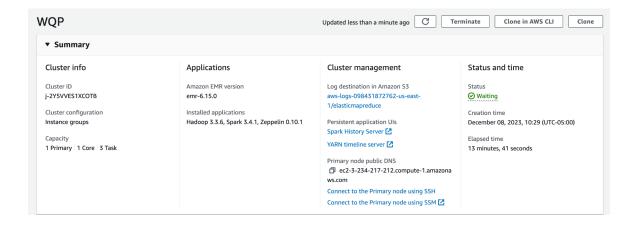
Create a key pair for SSH and SCP to the Instance/Cluster:



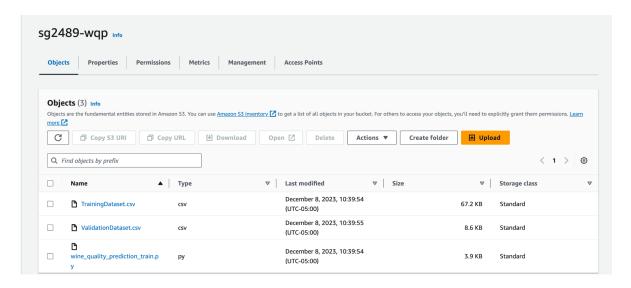
Select default IAM roles:



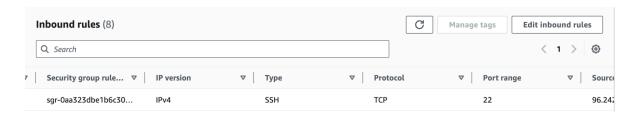
Rest, go with default configuration, cluster/s will be working upon turning to 'Waiting' status.



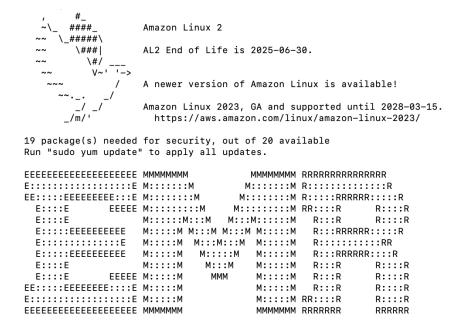
Create an S3 Bucket and drop training and testing files into it.



Ensure master instance of the cluster has permission to SSH, if not edit inbound rule after navigating to the security group of the instance.



Done SSH into my master instance using below command: ssh -i "<my-key-pair.pem>" ec2-user@<Public DNS>



Navigated to root user using: sudo su

Now, trained the model with training file using the below command: spark-submit s3://
bucket-name>/wine_quality_prediction_train.py

After successful training, results of trained model are:

```
Test Accuracy of Initial Model: 0.99375
/usr/lib/spark/python/lib/pyspark.zip/pyspark/sql/context
Weighted f1 score of Initial Model: 0.9933730158730157
Best Model: PipelineModel_55a5c603be56
Test Accuracy of Best Model: 0.96875
Weighted f1 score of Best Model: 0.954791666666667
```

Running Single Machine App in EC2 without Docker

Fetch trained model from S3 to master instance into a new folder (trained_model) using below command:

aws s3 sync aws s3 sync s3://<my-bucket-name>/trained_model.model ./trained_model

Done SCP for <testdataset> and Single Machine Prediction code file using below command: scp -i <keypair.pem> <file> ec2-user@<Public DNS>:/home/ec2-user

Installed Pyspark with: pip install pyspark

FutureWarning,
Weighted F1 score of the Wine Quality Prediction Model: 0.9779339666913879

for running single machine prediction application with external test dataset, using below command: python wqp_single_machine.py

probability pred	iction					,	lfur dioxide density pH sul				Label rawPrediction
8.9		0.22	0.48	1.8	0.077	29.0	60.0 0.9968 3.39		9.4		1.0 [19.4518621405359 [0.0
372428107 7.6	1.0	0.39	0.31	2.3	0.082	23.0	71.0 0.9982 3.52	0.65	9.7	5.0 [7.6,0.39,0.31,2	0.0 [479.531456024878 [0.9
7.9	0.0	0.43	0.21	1.6	0.106	10.0	37.0 0.9966 3.17	0.91	9.5	5.0 [7.9,0.43,0.21,1	0.0 [484.885809101067 [0.9
8.5 841896582	0.0	0.49	0.11	2.3	0.084	9.0	67.0 0.9968 3.17	0.53	9.4	5.0 [8.5,0.49,0.11,2	0.0 [491.163209482913 [0.9
6.9	1.0	0.4	0.14	2.4	0.085	21.0	40.0 0.9968 3.43	0.63	9.7	6.0 [6.9,0.4,0.14,2.4	1.0 [3.29191168480642 [0.0

Running Single Machine App in Local with Docker:

Fetch trained model from master instance into local (in a new folder wine_quality_predictor) using below command:

scp -i <keypair.pem> -r ec2-user@<Public DNS>:/home/ec2-user/trained_model ./wine_quality_predictor

Install and sign up into Docker Desktop, login to Docker from terminal with command: 'docker login'

cd to the source folder and build the image of the app: 'docker build -t <image_name:version> .'

run the app: 'docker run <image name:version>'

/usr/local/lib/python3.7/site-packages/pyspark/sql/context.py:159: FutureWarning: Deprecated in 3.0.0. Use SparkSession.builder.getOrCreate() instead.

		ty citric acid	residual sugar c	hlorides free	sulfur dioxide total sul	lfur dioxide density pH su	lphates alco	hol quality	features 1	abel rawPrediction
probability predi		+						+	+	
	+									
7.4		.7 0.0	1.9	0.076	11.0	34.0 0.9978 3.51	0.56	9.4 5.0	[7.4,0.7,0.0,1.9,	0.0 [488.247586263641 [0.9
517252728	0.0									
7.8	0.8	38 0.0	2.6	0.098	25.0	67.0 0.9968 3.2	0.68	9.8 5.0	[7.8,0.88,0.0,2.6]	0.0 [476.049269864772 [0.9
353972954	0.0			0.0001	45.01	5/ 01 0 00710 0/1	0.451		117 0 0 7/ 0 0/ 0	0.015/3/ 00404335//0/ 150.6
7.8	0.7	76 0.04	2.3	0.092	15.0	54.0 0.997 3.26	0.65	9.8 5.0	[7.8,0.76,0.04,2	0.0 [474.901917756696 [0.9
383551339	0.0	281 0.561	1.91	0.075	17.0	60.0 0.998 3.16	0.581	9.81 6.0	IF11 2 0 20 0 E4 1	1.0 [26.3858382129224 [0.0
167642584	1.01	201 0.001	1.7	0.0/5	17.0	00.0 0.770 3.10	0.001	7.01 0.0	[[11.2,0.20,0.30,1]	1.0 [20.3030302129224 [0.0
7.4		.71 0.01	1.91	0.076	11.0	34.0 0.9978 3.51	0.561	9.4 5.0	174979919	0.0 [488.247586263641 [0.9
517252728	0.01	.,,,	2171	010/01	11101	0410 017770 0102	01001	7141 010	10114/011/010/21//1111	010 [1-0012-7-0002000-21111 [017
								+		

None Test Accuracy of the Wine Quality Prediction Model: 8.96875 Weighted F1 score of the Wine Quality Prediction Model: 8.954_7916666666667 run the app with external dataset: docker run -v /Users/<username>/<path_to_dataset_directory>:/<path_on_host> <image_name:version> <external_dataset>

Sample Predictions	::	+							+	
	+									
		idity cit	ric acid residua	l sugar ch	lorides free	sulfur dioxide total sulfu	r dioxide density pH sul	phates alo	ohol qu	ality features label rawPrediction
probability pred										
		+								
8.91		0.22	0.48	1.8	0.077	29.0	60.0 0.9968 3.39	0.53	9.4	6.0 [8.9,0.22,0.48,1 1.0 [19.4518621405359 [0.0389
0372428107	1.0					·				
7.6		0.39	0.31	2.3	0.082	23.0	71.0 0.9982 3.52	0.65	9.7	5.0 [7.6,0.39,0.31,2 0.0 [479.531456024878 [0.9590
6291204975	0.0									
7.9		0.43	0.21	1.6	0.106	10.0	37.0 0.9966 3.17	0.91	9.5	5.0 [7.9,0.43,0.21,1 0.0 [484.885809101067 [0.9697
7161820213	0.0									
8.5		0.49	0.11	2.3	0.084	9.0	67.0 0.9968 3.17	0.53	9.4	5.0 [8.5,0.49,0.11,2 0.0 [491.163209482913 [0.9823
2641896582	0.0									
6.9		0.4	0.14	2.4	0.085	21.0	40.0 0.9968 3.43	0.63	9.7	6.0 [6.9,0.4,0.14,2.4 1.0 [3.29191168480642 [0.0065
8382336961	1.0									
+		+			+				+	
	+									

only showing top 5 rows

None Test Accuracy of the Wine Quality Prediction Model: 0.9843627834245594 Weighted F1 score of the Wine Quality Prediction Model: 0.9779339666913879

Create a new repo in DockerHub.

Push the image into repo with the below commands: docker tag <img:ver> <user_id>/<repo>:<tag> docker push <user_id>/<repo>:<tag>

Running Single Machine App in EC2 with Docker:

Create an EC2 instance with default configurations

Install docker with command: sudo yum install docker

Start docker: sudo service docker start

Pull the image into EC2: sudo docker pull <user_id>/<repo>:<tag>

Run the app: sudo docker run <user id>/<repo>:<tag>

Sample Predictions:								
·		·						
		citric acid resid	lual sugar∣chlori	des free sulfur dioxide tota	al sulfur dioxide density p	H sulphates a]	cohol qu	uality features label rawPrediction
probability predi				+				
	+							
8.9	0.22	0.48	1.8 0.	977 29.0	60.0 0.9968 3.3	9 0.53	9.4	6.0 [8.9,0.22,0.48,1 1.0 [19.4518621405359 [0.0389
0372428107	1.0							
7.6	0.39	0.31	2.3 0.	982 23.0	71.0 0.9982 3.5	2 0.65	9.7	5.0 [7.6,0.39,0.31,2 0.0 [479.531456024878 [0.9590
6291204975	0.0							
7.9	0.43	0.21	1.6 0.	106 10.0	37.0 0.9966 3.1	7 0.91	9.5	5.0 [7.9,0.43,0.21,1 0.0 [484.885809101067 [0.969]
7161820213	0.0							
8.5	0.49	0.11	2.3 0.	9.0	67.0 0.9968 3.1	7 0.53	9.4	5.0 [8.5,0.49,0.11,2 0.0 [491.163209482913 [0.982
2641896582	0.0							
6.9	0.4	0.14	2.4 0.	985 21.0	40.0 0.9968 3.4	3 0.63	9.7	6.0 [6.9,0.4,0.14,2.4 1.0 [3.29191168480642 [0.006
8382336961	1.0							

only showing top 5 rows

None Test Accuracy of the Wine Quality Prediction Model: 0.9843627834245594 Weighted F1 score of the Wine Quality Prediction Model: 0.9779339666913879