

Name: Abishek Vellineni

Email Id: av739@njit.edu

Module 07 Assignment 01: Programming Assignment 2

Wine Quality Prediction AWS Spark Application

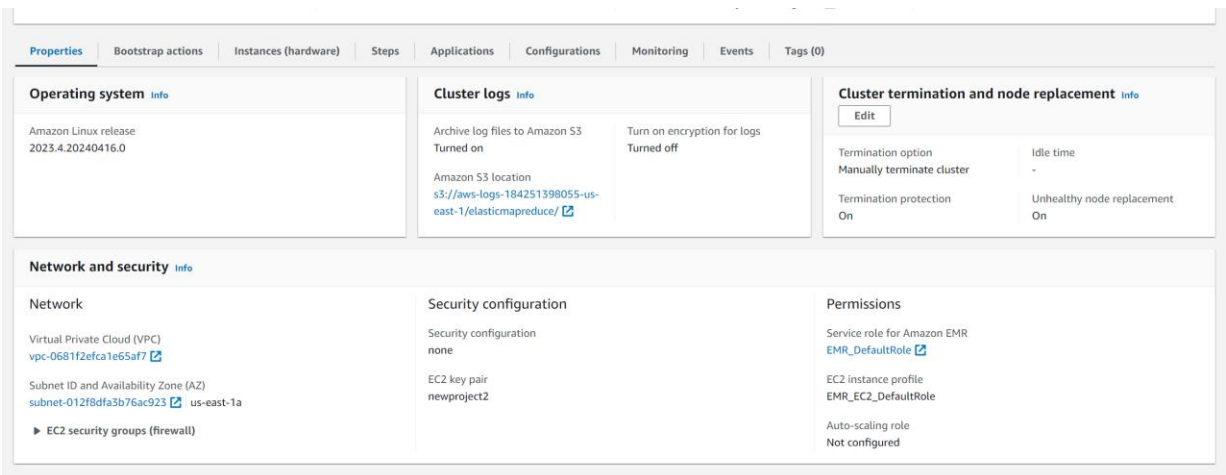
The application is deployed on an Amazon Web Services (AWS) Elastic MapReduce (EMR) cluster. The primary objective is to parallelly train a machine learning model on EC2 instances for predicting wine quality using publicly available data. Subsequently, the trained model is employed to predict the quality of wine. Docker is utilized to create a container image for the trained machine learning model, streamlining the deployment process.

Link for GitHub - https://github.com/Abishek183/wine_prediction

Link for Docker - https://hub.docker.com/repository/docker/abishek183/wine_predict/tags

Cluster Creation AWS SPARK:

1. Navigate to the EMR console and then click on create new instance.
2. Provide the name for your cluster.
3. In cluster Termination, change to manual from automatic.
4. Provide the key_pair which is a .pem file in security configuration



5. In instance creation, provide 1 for core and 4 for tasks as we need to run on 4 EC2 instances.
6. Then select default roles for the IAM roles.

Properties | Bootstrap actions | **Instances (hardware)** | Steps | Applications | Configurations | Monitoring | Events | Tags (0)

Instance group settings info Edit cluster scaling option

Cluster scaling option
Manually set cluster size

Core
Name and maximum core nodes in the cluster
Core: 1 instances

Task
Name and maximum task nodes in the cluster
Task: 1 - 4 instances

Instance groups (3) info Terminate instance Resize instance group Add task instance group

With the instance groups configuration, each node type consists of the same instance type and the same purchasing option for instances: on-demand or spot.

Find instances by status Find resources by ID or type; or search for text within loaded results

Type and name	ID	Status Last state change reason	Instances	Purchasing option and p...	EBS size (GiB)	EC2 instance ID
Primary	ig-21YXG8XL97N	Running	1	On-demand	-	-
Core	ig-27S23Z7KOU018	Running	1	On-demand	-	-
Task (Task - 1)	ig-3OUBUOCQM2B6	Running	4	On-demand	-	-

- Create a S3 queue to upload the python and csv files.
s3://winepredcit

aws Services Search [Alt+S] N. Virginia voclabs/user3025972-av/739@njit.edu @ 1842-5139-8055

Amazon S3 > Buckets > winepredcit

winepredcit info

Objects | Properties | Permissions | Metrics | Management | Access Points

Objects (5) 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

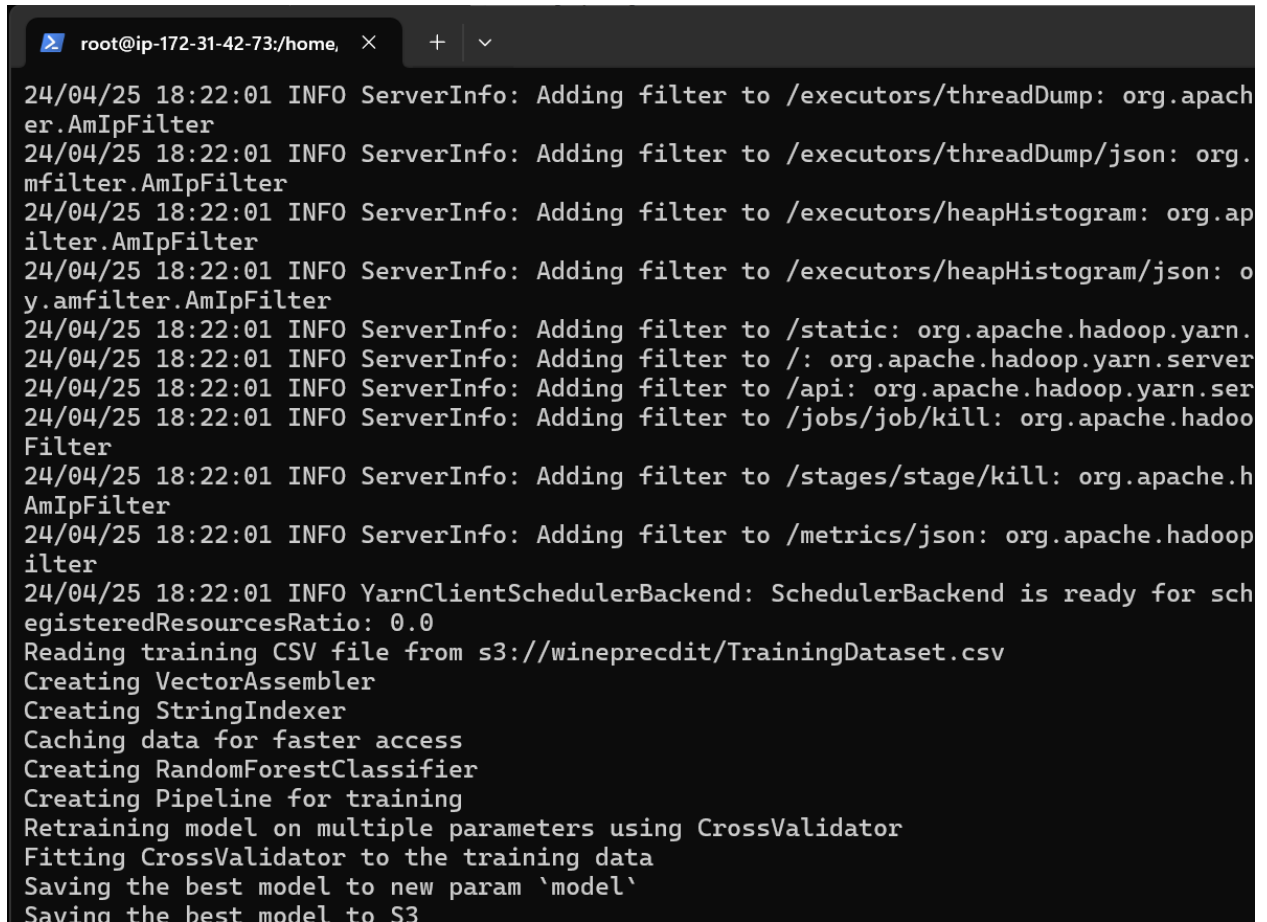
Name	Type	Last modified	Size	Storage class
prediction.py	py	April 25, 2024, 13:57:56 (UTC-04:00)	2.0 KB	Standard
trainedmodel/	Folder	-	-	-
training.py	py	April 25, 2024, 13:57:57 (UTC-04:00)	3.5 KB	Standard
TrainingDataset.csv	csv	April 24, 2024, 20:57:07 (UTC-04:00)	67.2 KB	Standard
ValidationDataset.csv	csv	April 24, 2024, 20:57:07 (UTC-04:00)	8.6 KB	Standard

CloudShell Feedback © 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

- Open terminal and use the below command to connect to the cluster.
ssh -i ~/newproject2.pem hadoop@ec2-52-201-250-228.compute-1.amazonaws.com

Execution without Docker

- Run “sudo su” command to change user.
- Install numpy by using ‘pip install numpy --user’
- Then run “spark-submit s3://wineprecidit/training.py”. It runs the file from S3 bucket and creates a ML model

A terminal window with a dark background and light-colored text. The window title bar shows 'root@ip-172-31-42-73:/home,'. The terminal output consists of multiple lines of log messages. The first part shows several 'INFO ServerInfo: Adding filter to /executors/...' messages for various components like threadDump, json, heapHistogram, and static. This is followed by 'INFO YarnClientSchedulerBackend: SchedulerBackend is ready for sch' and 'egisteredResourcesRatio: 0.0'. The main part of the log describes the training process: 'Reading training CSV file from s3://wineprecidit/TrainingDataset.csv', 'Creating VectorAssembler', 'Creating StringIndexer', 'Caching data for faster access', 'Creating RandomForestClassifier', 'Creating Pipeline for training', 'Retraining model on multiple parameters using CrossValidator', 'Fitting CrossValidator to the training data', 'Saving the best model to new param `model`', and 'Saving the best model to S3'.

- Then run “spark-submit s3://wineprecidit/prediction.py s3://wineprecidit/ValidationDataset.csv”. It uses the model created and validates the data from the csv file and provides the result.
- We can infer that it provides an accuracy of 95.4% from the below image.


```
root@ip-172-31-42-73/home, x + v
mIpFilter
24/04/25 18:24:08 INFO ServerInfo: Adding filter to /executors: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/25 18:24:08 INFO ServerInfo: Adding filter to /executors/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/25 18:24:08 INFO ServerInfo: Adding filter to /executors/threadDump: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/25 18:24:08 INFO ServerInfo: Adding filter to /executors/threadDump/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/25 18:24:08 INFO ServerInfo: Adding filter to /executors/heapHistogram: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/25 18:24:08 INFO ServerInfo: Adding filter to /executors/heapHistogram/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/25 18:24:08 INFO ServerInfo: Adding filter to /static: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/25 18:24:08 INFO ServerInfo: Adding filter to /: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/25 18:24:08 INFO ServerInfo: Adding filter to /api: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/25 18:24:08 INFO ServerInfo: Adding filter to /jobs/job/kill: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/25 18:24:08 INFO ServerInfo: Adding filter to /stages/stage/kill: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/25 18:24:08 INFO ServerInfo: Adding filter to /metrics/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/25 18:24:08 INFO YarnClientSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.0
Test Accuracy of wine prediction model = 0.96875
/usr/lib/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
[root@ip-172-31-42-73 hadoop]#
```

Execution with Docker

- Run the below commands to start the docker in the EC2.
 - sudo systemctl start docker
 - sudo systemctl enable docker
- Get the image from docker repo using the below command.
 - sudo docker pull abishek183/wine_predict:train
 - sudo docker pull abishek183/wine_predict:predict
- Run the train tag image to create a ML model using the below command.
 - sudo docker run -v /home/ec2-user/:/job abishek183/wine_predict:train
- Run the image with predict tag to get the accuracy.
 - sudo docker run -v /home/ec2-user/:/job abishek183/wine_predict:predict ValidationDataset.csv
- we can infer from the below image the accuracy is 95%.

```
root@ip-172-31-42-73:/home, x + v
a93f/fetchFileTemp4538105940248813041.tmp has been previously copied to /tmp/spark-ff09e041-4b82-4d22-add2-4b8c18c094c4/
userFiles-c8c334e8-749e-4fba-b172-2050f3a6a93f/net.java.dev.jets3t_jets3t-0.9.0.jar
24/04/25 18:25:17 INFO Executor: Adding file:/tmp/spark-ff09e041-4b82-4d22-add2-4b8c18c094c4/userFiles-c8c334e8-749e-4fb
a-b172-2050f3a6a93f/net.java.dev.jets3t_jets3t-0.9.0.jar to class loader default
24/04/25 18:25:17 INFO Executor: Fetching spark://e58aecac3fc2:38395/jars/org.apache.httpcomponents_httpcore-4.2.5.jar w
ith timestamp 1714069514920
24/04/25 18:25:17 INFO Utils: Fetching spark://e58aecac3fc2:38395/jars/org.apache.httpcomponents_httpcore-4.2.5.jar to /
tmp/spark-ff09e041-4b82-4d22-add2-4b8c18c094c4/userFiles-c8c334e8-749e-4fba-b172-2050f3a6a93f/fetchFileTemp5600856328625
415611.tmp
24/04/25 18:25:17 INFO Utils: /tmp/spark-ff09e041-4b82-4d22-add2-4b8c18c094c4/userFiles-c8c334e8-749e-4fba-b172-2050f3a6
a93f/fetchFileTemp5600856328625415611.tmp has been previously copied to /tmp/spark-ff09e041-4b82-4d22-add2-4b8c18c094c4/
userFiles-c8c334e8-749e-4fba-b172-2050f3a6a93f/org.apache.httpcomponents_httpcore-4.2.5.jar
24/04/25 18:25:17 INFO Executor: Adding file:/tmp/spark-ff09e041-4b82-4d22-add2-4b8c18c094c4/userFiles-c8c334e8-749e-4fb
a-b172-2050f3a6a93f/org.apache.httpcomponents_httpcore-4.2.5.jar to class loader default
24/04/25 18:25:17 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on
port 34061.
24/04/25 18:25:17 INFO NettyBlockTransferService: Server created on e58aecac3fc2:34061
24/04/25 18:25:17 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication p
olicy
24/04/25 18:25:17 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, e58aecac3fc2, 34061, None)
24/04/25 18:25:17 INFO BlockManagerMasterEndpoint: Registering block manager e58aecac3fc2:34061 with 366.3 MiB RAM, Bloc
kManagerId(driver, e58aecac3fc2, 34061, None)
24/04/25 18:25:17 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, e58aecac3fc2, 34061, None)
24/04/25 18:25:17 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, e58aecac3fc2, 34061, None)
Test Accuracy of wine prediction model = 0.96875
/opt/spark/python/lib/pyspark.zip/pyspark/sql/context.py:160: FutureWarning: Deprecated in 3.0.0. Use SparkSession.build
er.getOrCreate() instead.
Weighted F1 Score of wine prediction model = 0.9541901629072682
Exiting Spark Application
[root@ip-172-31-42-73 hadoop]#
```

Docker Image:



abishek183/wine_predict:train

MANIFEST DIGEST sha256:0be9bde76284457b9b7a7b825f932c0f6de7f70246313c8dc2d804d89656

OS/ARCH linux/amd64COMPRESSED SIZE 819.25 MBLAST PUSHED an hour ago by [abishek183](#)TYPE ImageMANIFEST DIGEST sha256:0be9bde...

Image LayersVulnerabilities

IMAGE LAYERS

1	ADD file ... in /	72.57 MB	Command
2	LABEL org.label-schema.schema-version=1.0 org.label-s...	0 B	ADD file:b3ebbebdc04723e43b7b44ad099bc057b63d93da2e9293983c30bfc1efaf53 in /
3	CMD ["bin/bash"]	0 B	
4	RUN /bin/sh -c yum -y	212.55 MB	
5	RUN /bin/sh -c python3 -V	93 B	
6	ENV PYSPARK_DRIVER_PYTHON=python3	0 B	
7	ENV PYSPARK_PYTHON=python3	0 B	
8	RUN /bin/sh -c pip3 install	4.48 MB	
9	RUN /bin/sh -c pip3 install	26.77 MB	
10	RUN /bin/sh -c pip3 install	53.36 MB	
11	WORKDIR /opt	32 B	
12	RUN /bin/sh -c wget --no-verbose	382.41 MB	
13	RUN /bin/sh -c yum -y	52.13 MB	
14	ENV SPARK_HOME=/opt/spark	0 B	
15	ENV PATH=/opt/spark/bin:/usr/local/bin:/usr/local/...	0 B	
16	RUN /bin/sh -c wget https://repo.maven.org/m...	18.69 MB	
17	RUN /bin/sh -c wget https://repo.maven.org/m...	278.38 KB	
18	ADD training.py . # buildkit	1.35 KB	



abishek183/wine_predict:predict

MANIFEST DIGEST: sha256:4b8230c6c234506723f92678209434e5d20485d70e53a80241267ee0e9ba35

Delete Tag

OS/ARCH	COMPRESSED SIZE	LAST PUSHED	TYPE	MANIFEST DIGEST
linux/amd64	815.24 MB	2 hours ago by abishek183	Image	sha256:4b8230c6...

Image Layers Vulnerabilities

IMAGE LAYERS

1	ADD file ... in /	72.57 MB
2	LABEL org.label-schema.schema-version=1.0 org.label...	0 B
3	CMD ["/bin/bash"]	0 B
4	RUN /bin/sh -c yum -y	212.55 MB
5	RUN /bin/sh -c python3 -V	93 B
6	DWY PYSPARK_DRIVER_PYTHON-python3	0 B
7	DWY PYSPARK_PYTHON-python3	0 B
8	RUN /bin/sh -c pip3 install	4.48 MB
9	RUN /bin/sh -c pip3 install	26.77 MB
10	RUN /bin/sh -c pip3 install	53.36 MB
11	WORKDIR /opt	32 B
12	RUN /bin/sh -c wget --no-verbose	382.41 MB
13	RUN /bin/sh -c yum -y	52.13 MB
14	DWY SPARK_HOME=/opt/spark	0 B
15	DWY PATH=/opt/spark/bin:/usr/local/sbin:/usr/local/...	0 B
16	RUN /bin/sh -c wget https://repo.maven.org/maven...	18.69 MB
17	RUN /bin/sh -c wget https://repo.maven.org/maven...	278.38 KB
18	ADD prediction.py . # buildkit	1.82 KB

Command

ADD file:b34bbe8bd384723643b7b44ad098cd8576a3d938a2a2929393a30bfc10fa53 in /