

# Wine Quality program deployment process

Name: Keerthi Kondisetty

UCID: rkk3

Github with dockerfile - <https://github.com/keerthikondisetty/cs643-wineTaster>

## Deployment and preparation of the instances

1. Launch an EMR cluster to run spark on the AWS Console.

The screenshot shows the AWS EMR console with the cluster 'my-wine-cluster' selected. The 'Summary' tab is active, displaying basic cluster information: ID: J1H9K000AWT24UJB, Creation date: 2017-07-28 21:28 (UTC-4), Elapsed time: 0 seconds, After last step completes: Cluster waits, Termination protection: Off, Tags: View All / Edit, Master public DNS: --. The 'Configuration details' section shows the release label as emr-5.36.0, Hadoop distribution as Amazon Hadoop 2.7.1, Applications as Hive 2.3.5, Hue 4.10.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.9.2, Log URI as s3://aws-logs-36554294936-us-east-1, and EMRFS consistent view as Enabled. The 'Application user interfaces' section lists Persistent user interfaces (On-cluster user interface) and Custom AMI ID. The 'Network and hardware' section shows the availability zone as us-east-1, subnet ID as subnet-d9c7a6eef1b42, and EC2 instance role as EMR\_EC2\_DefaultRole. The 'Security and access' section shows the key name as id-12, EC2 instance profile as EMR\_EC2\_DefaultRole, and EMR role as EMR\_DefaultRole. The 'Core' provisioning status is shown as 1 master, 2 m3-large, and 1 task. Cluster scaling and auto-termination settings are also listed.

2. Using the command below, Copy the whole directory onto the master node.

```
scp -i ~/.ssh/id_rsa -r ${pwd}/*  
hadoop@<public_dns_master_node>:/home/master/wineTaster
```

3. Start the worker machines using the below command and use the next command to check if nodes are online.

1. Start-workers.sh
2. jps

```
ubuntu@Master: $ start-workers.sh
172.31.89.84: org.apache.spark.deploy.worker.Worker running as process 2847. Stop it first.
172.31.92.99: starting org.apache.spark.deploy.worker.Worker, logging to /home/ubuntu/cs643/spark-3.1.2-bin-hadoop3.2/logs/spark-ubuntu-org.apache.spark.deploy.worker.Worker-1-Master.out
localhost: starting org.apache.spark.deploy.worker.Worker, logging to /home/ubuntu/cs643/spark-3.1.2-bin-hadoop3.2/logs/spark-ubuntu-org.apache.spark.deploy.worker.Worker-1-Master.out
ubuntu@Master: $ jps
3296 Worker
3382 Jps
3490 Master
3275 Worker
ubuntu@Master: $
```

4. Start the training by running the file using “python training.py <Dataset>”.

*Before running the above command, make sure you are in the same directory as you copied the files in step 2.*

Running the actual prediction using the training model created above.

1. We will pull the docker image from the docker hub container registry

```
docker pull rk33/cs643-wineTaster
```

2. Run the docker image using the below command

```
docker container run -rm -it rk33/cs643-wineTaster
```

3. The result should be something similar to the below image.

```
02:54:40 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 3 to 172.31.4
02:54:40 INFO TaskSetManager: Finished task 0.0 in stage 126.0 (TID 125) in 18 ms on ip-172-31-49-119.ec2.
cutor 1) (1/1)
02:54:40 INFO YarnScheduler: Removed TaskSet 126.0, whose tasks have all completed, from pool
02:54:40 INFO DAGScheduler: ResultStage 126 (collectAsMap at MulticlassMetrics.scala:53) finished in 0.024
02:54:40 INFO DAGScheduler: Job 122 finished: collectAsMap at MulticlassMetrics.scala:53, took 0.114820 s
e: 0.5904050519731796
02:54:40 INFO SparkContext: Invoking stop() from shutdown hook
02:54:40 INFO SparkUI: Stopped Spark web UI at http://ip-172-31-60-95.ec2.internal:4040
02:54:40 INFO YarnClientSchedulerBackend: Interrupting monitor thread
02:54:40 INFO YarnClientSchedulerBackend: Shutting down all executors
02:54:40 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
02:54:40 INFO SchedulerExtensionServices: Stopping SchedulerExtensionServices
Option=None,
s>List(),
l=false)
02:54:40 INFO YarnClientSchedulerBackend: Stopped
02:54:40 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
02:54:40 INFO MemoryStore: MemoryStore cleared
02:54:40 INFO BlockManager: BlockManager stopped
02:54:40 INFO BlockManagerMaster: BlockManagerMaster stopped
02:54:40 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
02:54:40 INFO SparkContext: Successfully stopped SparkContext
02:54:40 INFO ShutdownHookManager: Shutdown hook called
02:54:40 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-38a9db92-1163-49b3-88bb-e37110fbe8c3
02:54:40 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-38a9db92-1163-49b3-88bb-e37110fbe8c3/
```