Contributor: Mohammed Hannan Desai(mmd76)

Wine Quality Prediction App using Apache Spark over AWS EMR Cluster

This project demonstrated the use of Apache Spark on AWS EMR Cluster to train machine learning model parallelly on multiple EC2 instance, 4 in our case and then save that model and use it to predict the quality of wine.

Data contains the information about wine quality, which is publicly available, so data has been available in format of two csv files TrainingDataSet.csv & ValidationDataSet.csv, both will be used to train and validate ML model.

This project contains two applications:

Both applications are developed using Java language and maven project management tool.

Link to download these two provided below:

1. **Application for build and train the model (wineprediction-model-training):** This application is used to train machine learning model using Linear Regression Algorithm and once model is created & trained, we will save it on Amazon s3 bucket.

   GitHub Repo URL:
   https://github.com/hannandesai/wineprediction-model-training.git

   This application will be run over Amazon EMR cluster using Apache Spark over 4 EC2 instance.

   Command to run this app:

   mvn clean package

   ```
   mvn exec:java –Dexec.mainClass="org.example.wineprediction.App"
   ```

2. **Prediction Application (winepredictionapp):** This application will use the model created by first application to predict quality of wines which will be provided as input from TestData.csv file.

   GitHub Repo URL:
   https://github.com/hannandesai/winepredictionapp.git

This application will be run over single EC2 instance and will run using docker created image.

Docker Hub URL for image:

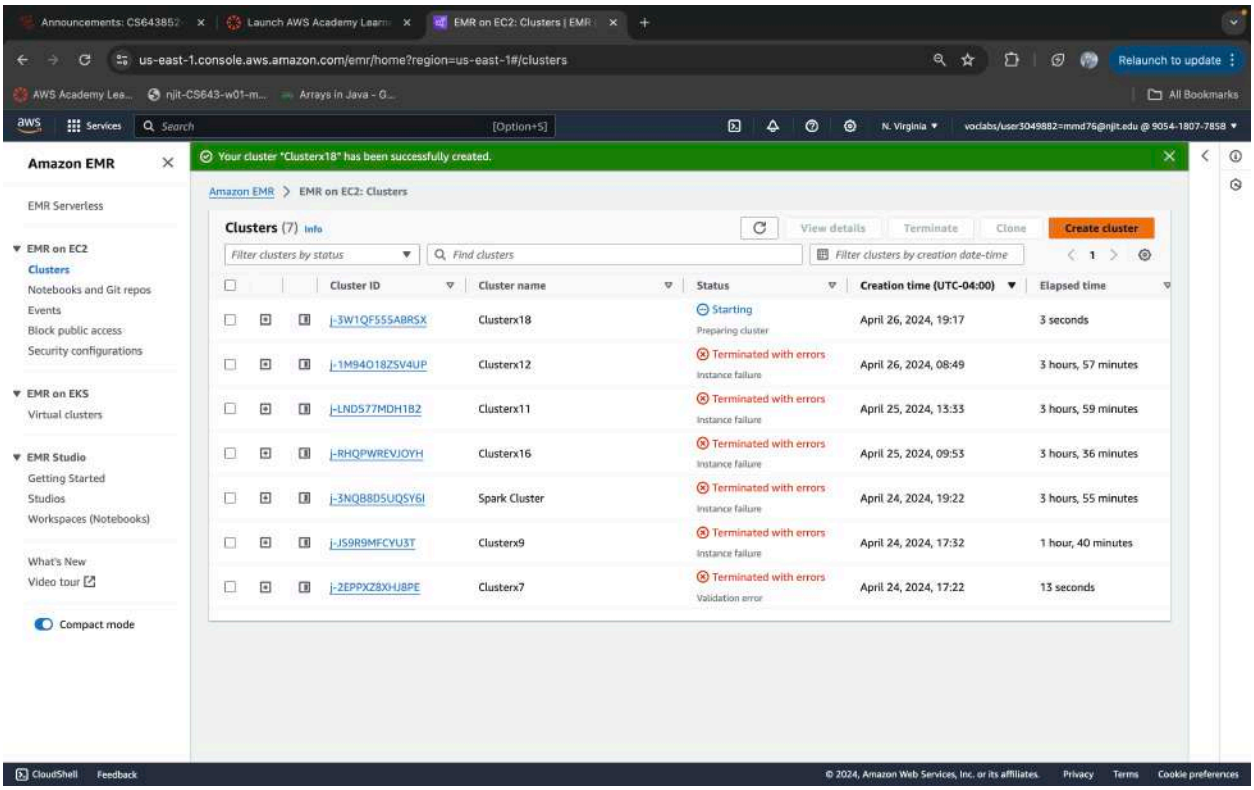
https://hub.docker.com/repository/docker/hannan9900/winepredictionapp/general

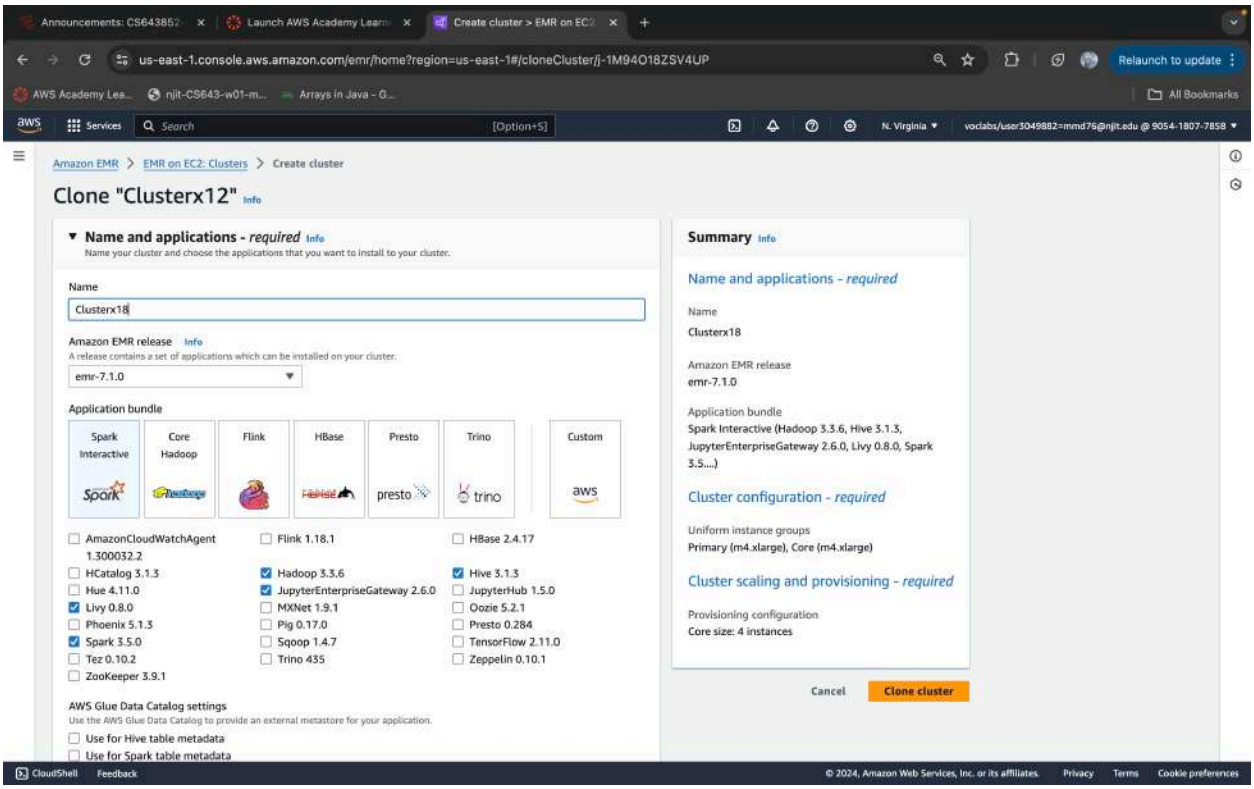
Command to run this app:

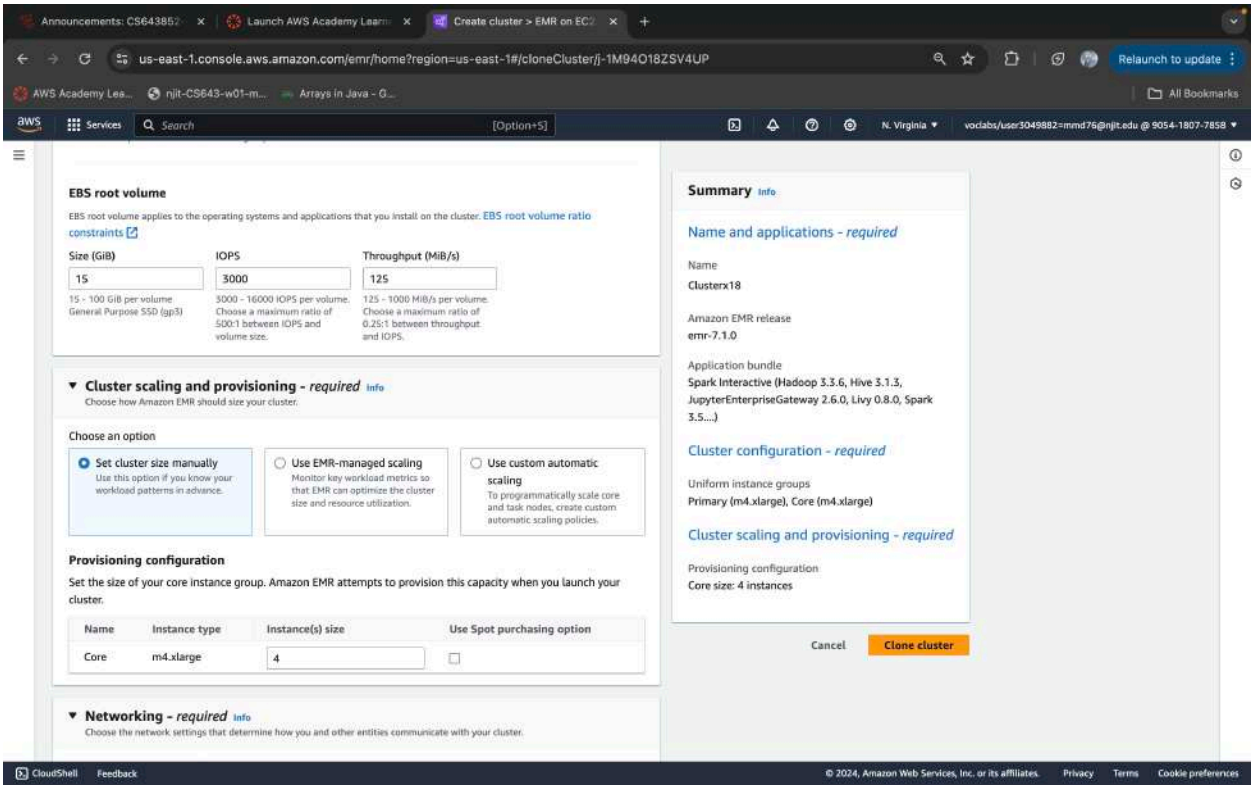
1. mvn clean package
2. `mvn exec:java -Dexec.mainClass="org.example.winepredictionapp.App"`

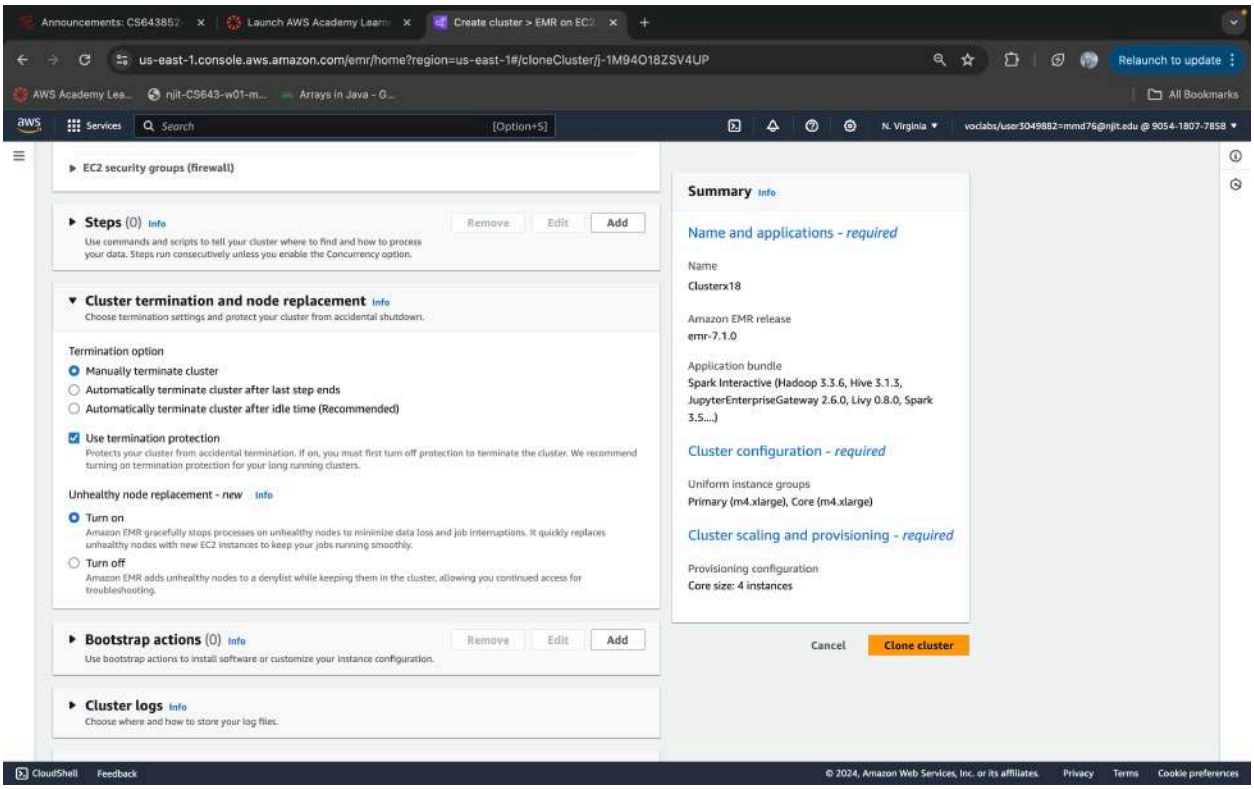
Cloud Setup:

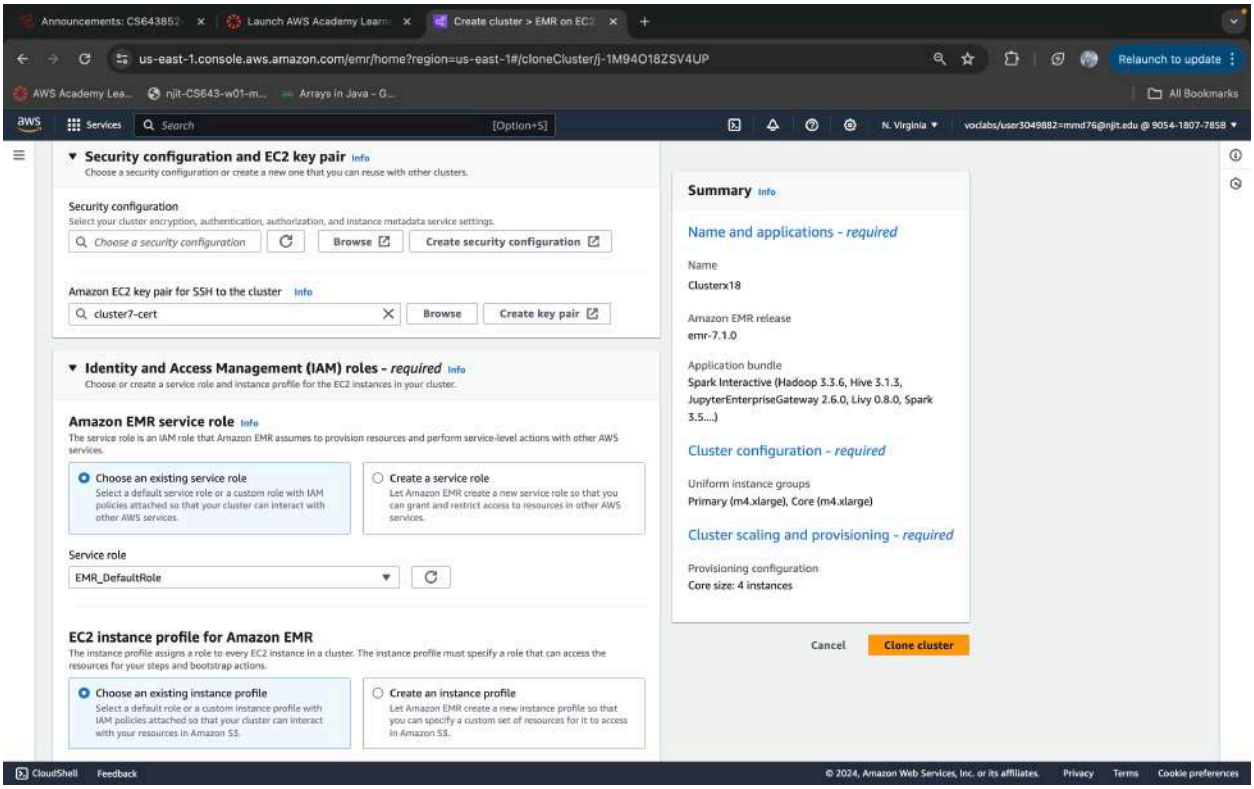
Setup EMR Cluster

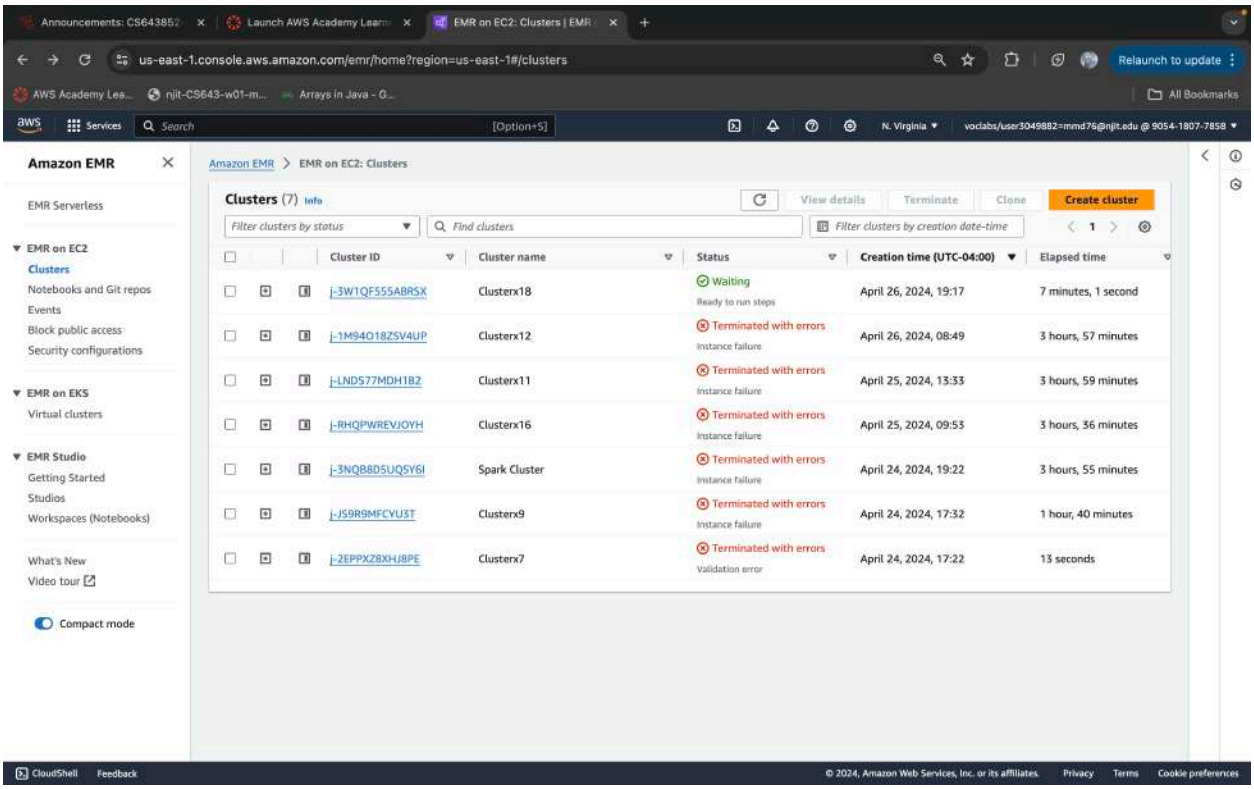


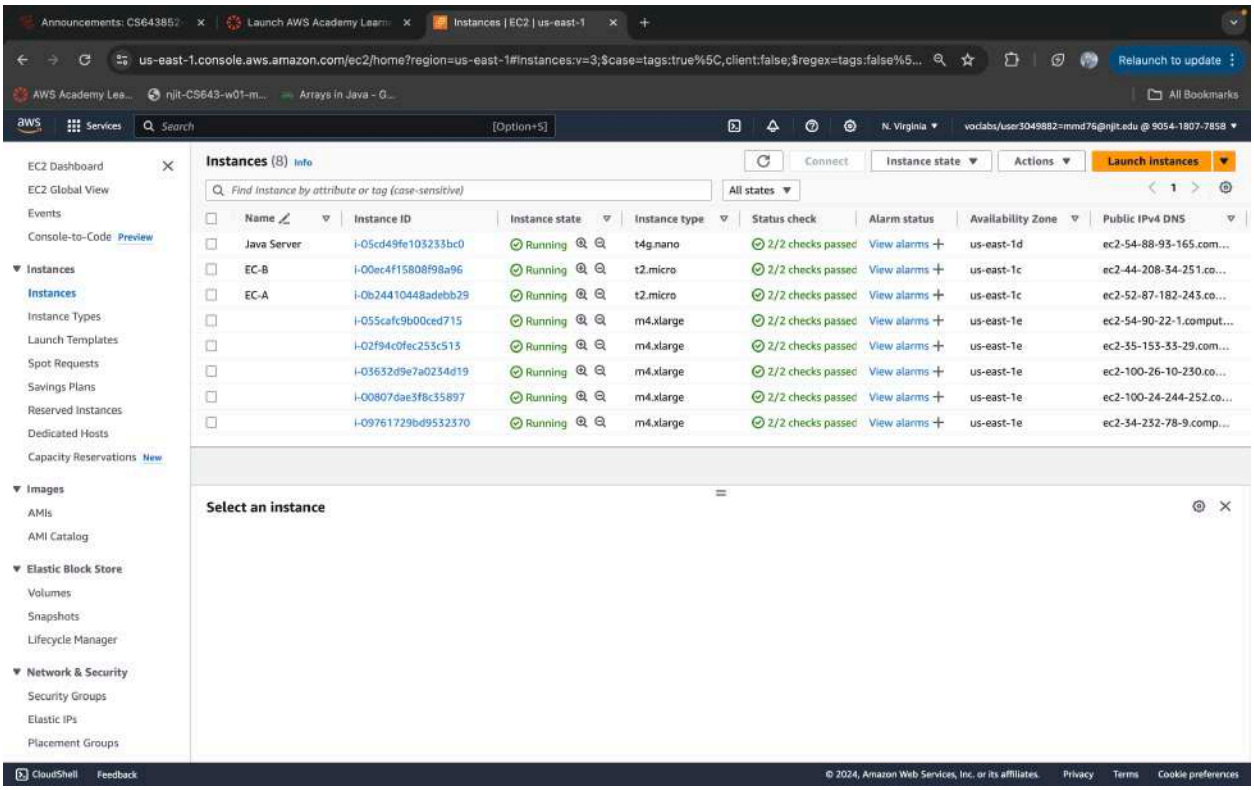

Announcements: CS643852
Launch AWS Academy Learn
Create cluster > EMR on EC2
us-east-1.console.aws.amazon.com/emr/home?region=us-east-1#/cloneCluster/j-1M94O18ZSV4UP
Relaunch to update
AWS Academy Lea...
njit-CS643-w01-m...
Arrays in Java - G...
All Bookmarks
Services
Search
[Option+S]
N. Virginia
voclabs/user3049882=mmd76@njit.edu @ 9054-1807-7858
Amazon EMR > EMR on EC2: Clusters > Create cluster
Clone "Clusterx12" Info
Name and applications - required Info
Name your cluster and choose the applications that you want to install to your cluster.
Name
Clusterx18
Amazon EMR release Info
A release contains a set of applications which can be installed on your cluster.
emr-7.1.0
Application bundle
Spark Interactive
Core Hadoop
Flink
HBase
Presto
Trino
Custom
AmazonCloudWatchAgent 1.300032.2
Flink 1.18.1
HBase 2.4.17
HCatalog 3.1.3
Hadoop 3.3.6
Hive 3.1.3
Hue 4.11.0
JupyterEnterpriseGateway 2.6.0
JupyterHub 1.5.0
Livy 0.8.0
MXNet 1.9.1
Oozie 5.2.1
Phoenix 5.1.3
Pig 0.17.0
Presto 0.284
Spark 3.5.0
Sqoop 1.4.7
TensorFlow 2.11.0
Tez 0.10.2
Trino 435
Zeppelin 0.10.1
ZooKeeper 3.9.1
AWS Glue Data Catalog settings
Use the AWS Glue Data Catalog to provide an external metastore for your application.
Use for Hive table metadata
Use for Spark table metadata
Summary Info
Name and applications - required
Name
Clusterx18
Amazon EMR release
emr-7.1.0
Application bundle
Spark Interactive (Hadoop 3.3.6, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.8.0, Spark 3.5....)
Cluster configuration - required
Uniform instance groups
Primary (m4.xlarge), Core (m4.xlarge)
Cluster scaling and provisioning - required
Provisioning configuration
Core size: 4 instances
Cancel
Clone cluster
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


Announcements: CS643852
Launch AWS Academy Learn
Create cluster > EMR on EC2
us-east-1.console.aws.amazon.com/emr/home?region=us-east-1#/cloneCluster/j-1M94O18ZSV4UP
Relaunch to update
AWS Academy Lea...
njit-CS643-w01-m...
Arrays in Java - G...
All Bookmarks
Services
Search
[Option+S]
N. Virginia
voclabs/user3049882=mmd76@njit.edu @ 9054-1807-7858
EBS root volume
EBS root volume applies to the operating systems and applications that you install on the cluster. EBS root volume ratio constraints
Size (GiB)
15
15 - 100 GiB per volume General Purpose SSD (gp3)
IOPS
3000
3000 - 16000 IOPS per volume. Choose a maximum ratio of 500:1 between IOPS and volume size.
Throughput (MiB/s)
125
125 - 1000 MiB/s per volume. Choose a maximum ratio of 0.25:1 between throughput and IOPS.
Cluster scaling and provisioning - required Info
Choose how Amazon EMR should size your cluster.
Choose an option
Set cluster size manually
Use this option if you know your workload patterns in advance.
Use EMR-managed scaling
Monitor key workload metrics so that EMR can optimize the cluster size and resource utilization.
Use custom automatic scaling
To programmatically scale core and task nodes, create custom automatic scaling policies.
Provisioning configuration
Set the size of your core instance group. Amazon EMR attempts to provision this capacity when you launch your cluster.
Name
Instance type
Instance(s) size
Use Spot purchasing option
Core
m4.xlarge
4
Networking - required Info
Choose the network settings that determine how you and other entities communicate with your cluster.
Summary Info
Name and applications - required
Name
Clusterx18
Amazon EMR release
emr-7.1.0
Application bundle
Spark Interactive (Hadoop 3.3.6, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.8.0, Spark 3.5....)
Cluster configuration - required
Uniform instance groups
Primary (m4.xlarge), Core (m4.xlarge)
Cluster scaling and provisioning - required
Provisioning configuration
Core size: 4 instances
Cancel
Clone cluster
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

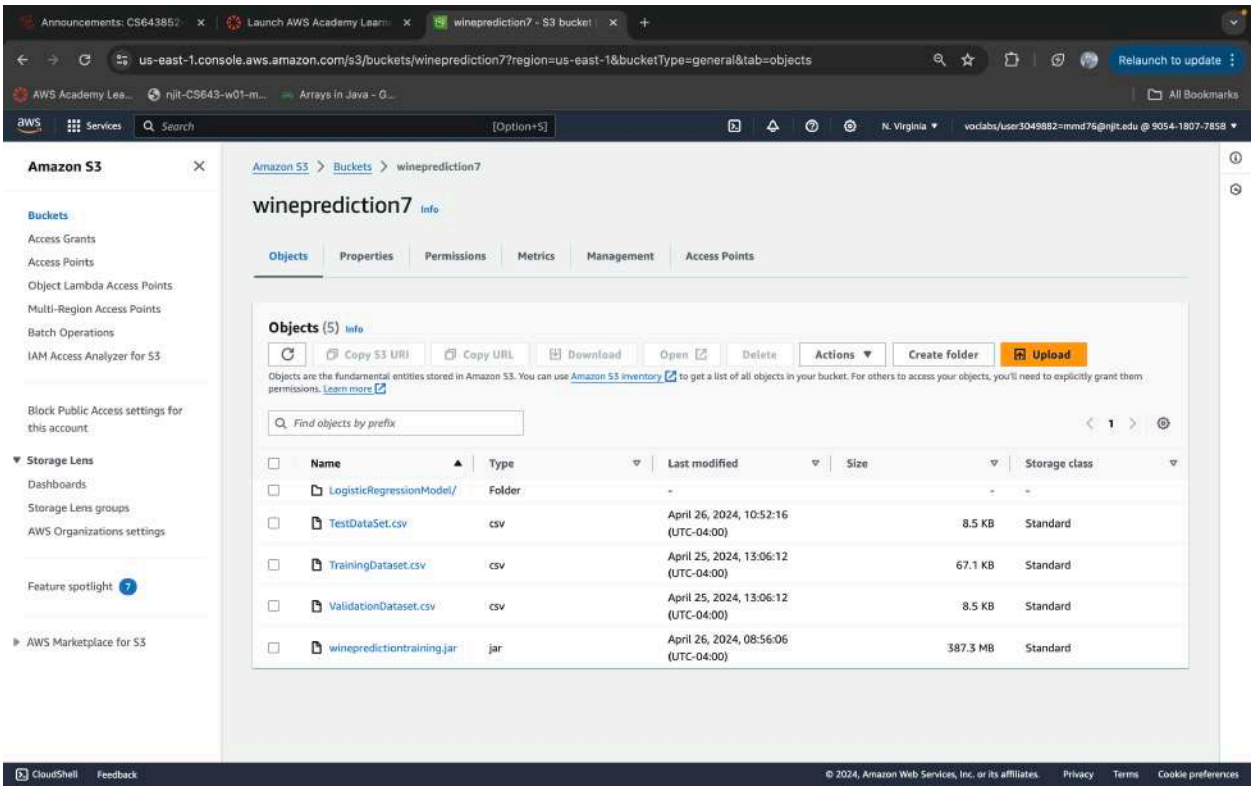

Announcements: CS643852
Launch AWS Academy Learn
Create cluster > EMR on EC2
us-east-1.console.aws.amazon.com/emr/home?region=us-east-1#/cloneCluster/j-1M94O18ZSV4UP
Relaunch to update
AWS Academy Lea...
njit-CS643-w01-m...
Arrays in Java - G...
All Bookmarks
Services
Search
[Option+S]
N. Virginia
voclabs/user3049882=mmd76@njit.edu @ 9054-1807-7858
EC2 security groups (firewall)
Steps (0) Info
Use commands and scripts to tell your cluster where to find and how to process your data. Steps run consecutively unless you enable the Concurrency option.
Remove
Edit
Add
Cluster termination and node replacement Info
Choose termination settings and protect your cluster from accidental shutdown.
Termination option
Manually terminate cluster
Automatically terminate cluster after last step ends
Automatically terminate cluster after idle time (Recommended)
Use termination protection
Protects your cluster from accidental termination. If on, you must first turn off protection to terminate the cluster. We recommend turning on termination protection for your long running clusters.
Unhealthy node replacement - new Info
Turn on
Amazon EMR gracefully stops processes on unhealthy nodes to minimize data loss and job interruptions. It quickly replaces unhealthy nodes with new EC2 instances to keep your jobs running smoothly.
Turn off
Amazon EMR adds unhealthy nodes to a denylist while keeping them in the cluster, allowing you continued access for troubleshooting.
Bootstrap actions (0) Info
Use bootstrap actions to install software or customize your instance configuration.
Cluster logs Info
Choose where and how to store your log files.
Summary Info
Name and applications - required
Name
Clusterx18
Amazon EMR release
emr-7.1.0
Application bundle
Spark Interactive (Hadoop 3.3.6, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.8.0, Spark 3.5....)
Cluster configuration - required
Uniform instance groups
Primary (m4.xlarge), Core (m4.xlarge)
Cluster scaling and provisioning - required
Provisioning configuration
Core size: 4 instances
Cancel
Clone cluster
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


Announcements: CS643852
Launch AWS Academy Learn
Create cluster > EMR on EC2
us-east-1.console.aws.amazon.com/emr/home?region=us-east-1#/cloneCluster/j-1M94O18ZSV4UP
Relaunch to update
AWS Academy Lea...
njit-CS643-w01-m...
Arrays in Java - G...
All Bookmarks
Services
Search
[Option+S]
N. Virginia
voclabs/user3049882=mmd76@njit.edu @ 9054-1807-7858
Security configuration and EC2 key pair Info
Choose a security configuration or create a new one that you can reuse with other clusters.
Security configuration
Select your cluster encryption, authentication, authorization, and instance metadata service settings.
Choose a security configuration
Browse
Create security configuration
Amazon EC2 key pair for SSH to the cluster Info
cluster7-cert
Browse
Create key pair
Identity and Access Management (IAM) roles - required Info
Choose or create a service role and instance profile for the EC2 instances in your cluster.
Amazon EMR service role Info
The service role is an IAM role that Amazon EMR assumes to provision resources and perform service-level actions with other AWS services.
Choose an existing service role
Select a default service role or a custom role with IAM policies attached so that your cluster can interact with other AWS services.
Create a service role
Let Amazon EMR create a new service role so that you can grant and restrict access to resources in other AWS services.
Service role
EMR_DefaultRole
EC2 instance profile for Amazon EMR
The instance profile assigns a role to every EC2 instance in a cluster. The instance profile must specify a role that can access the resources for your steps and bootstrap actions.
Choose an existing instance profile
Select a default role or a custom instance profile with IAM policies attached so that your cluster can interact with your resources in Amazon S3.
Create an instance profile
Let Amazon EMR create a new instance profile so that you can specify a custom set of resources for it to access in Amazon S3.
Summary Info
Name and applications - required
Name
Clusterx18
Amazon EMR release
emr-7.1.0
Application bundle
Spark Interactive (Hadoop 3.3.6, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.8.0, Spark 3.5....)
Cluster configuration - required
Uniform instance groups
Primary (m4.xlarge), Core (m4.xlarge)
Cluster scaling and provisioning - required
Provisioning configuration
Core size: 4 instances
Cancel
Clone cluster
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


Announcements: CS643852
Launch AWS Academy Learn
EMR on EC2: Clusters | EMR
us-east-1.console.aws.amazon.com/emr/home?region=us-east-1#/clusters
Relaunch to update
AWS Academy Lea...
njit-CS643-w01-m...
Arrays in Java - G...
All Bookmarks
Services
Search
[Option+S]
N. Virginia
voclabs/user3049882=mmd76@njit.edu @ 9054-1807-7858
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 (7) Info
View details
Terminate
Clone
Create cluster
Filter clusters by status
Find clusters
Filter clusters by creation date-time
Cluster ID
Cluster name
Status
Creation time (UTC-04:00)
Elapsed time
j-3W1QF555A8RSX
Clusterx18
Waiting
Ready to run steps
April 26, 2024, 19:17
7 minutes, 1 second
j-1M94O18ZSV4UP
Clusterx12
Terminated with errors
Instance failure
April 26, 2024, 08:49
3 hours, 57 minutes
j-LND577MDH1BZ
Clusterx11
Terminated with errors
Instance failure
April 25, 2024, 13:33
3 hours, 59 minutes
j-RHQPWREVJOYH
Clusterx16
Terminated with errors
Instance failure
April 25, 2024, 09:53
3 hours, 36 minutes
j-3NQB8D5UQSY6I
Spark Cluster
Terminated with errors
Instance failure
April 24, 2024, 19:22
3 hours, 55 minutes
j-JS9R9MFCVU3T
Clusterx9
Terminated with errors
Instance failure
April 24, 2024, 17:32
1 hour, 40 minutes
j-2EPPXZ8XHJ8PE
Clusterx7
Terminated with errors
Validation error
April 24, 2024, 17:22
13 seconds
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

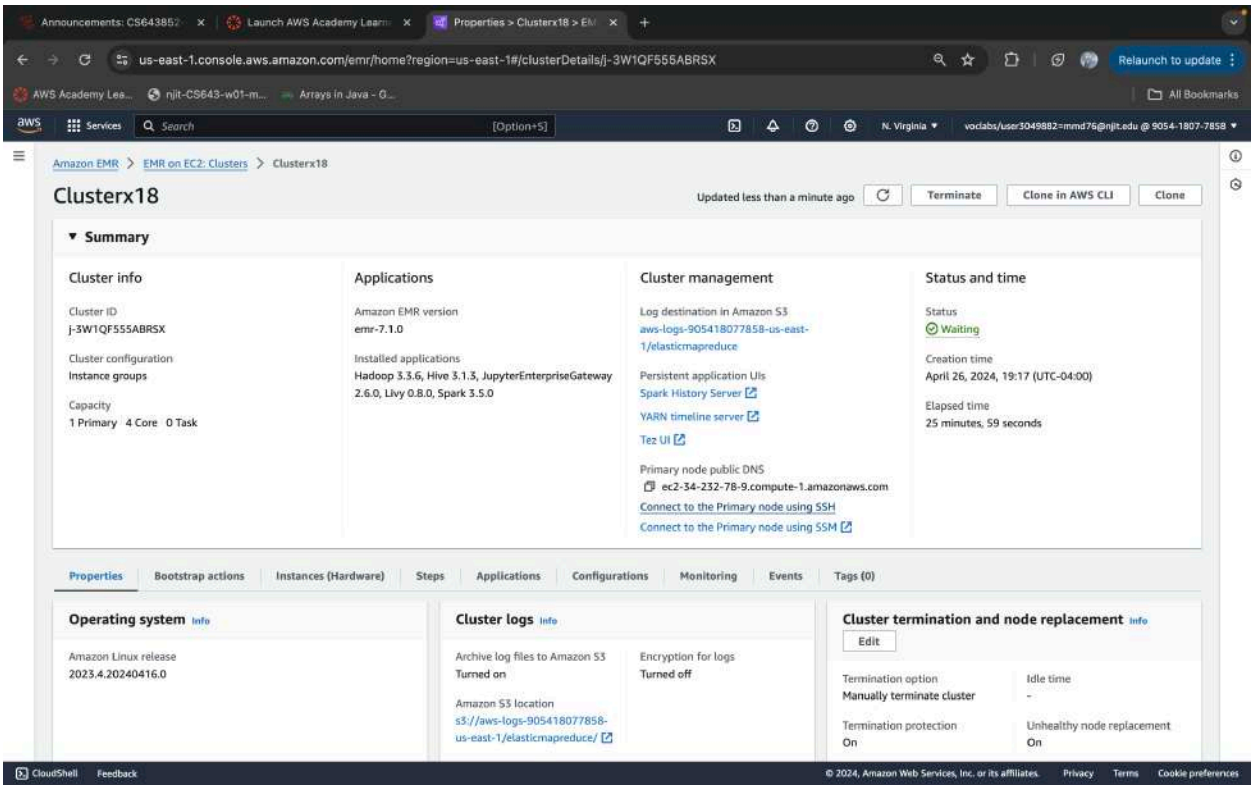

Announcements: CS643852
Launch AWS Academy Learn
Instances | EC2 | us-east-1
us-east-1.console.aws.amazon.com/ec2/home?region=us-east-1#instances:v=3;$case=tags:true%5C,client:false;$regex=tags:false%5...
Relaunch to update
AWS Academy Lea...
njit-CS643-w01-m...
Arrays in Java - G...
All Bookmarks
Services
Search
[Option+S]
N. Virginia
voclabs/user3049882=mmd76@njit.edu @ 9054-1807-7858
EC2 Dashboard
EC2 Global View
Events
Console-to-Code Preview
Instances
Instances
Instance Types
Launch Templates
Spot Requests
Savings Plans
Reserved Instances
Dedicated Hosts
Capacity Reservations New
Images
AMIs
AMI Catalog
Elastic Block Store
Volumes
Snapshots
Lifecycle Manager
Network & Security
Security Groups
Elastic IPs
Placement Groups
Instances (8) Info
Connect
Instance state
Actions
Launch instances
Find instance by attribute or tag (case-sensitive)
All states
Name
Instance ID
Instance state
Instance type
Status check
Alarm status
Availability Zone
Public IPv4 DNS
Java Server
i-05cd49fe103233bc0
Running
t4g.nano
2/2 checks passed
View alarms
us-east-1d
ec2-54-88-93-165.com...
EC-B
i-00ec4f15808f98a96
Running
t2.micro
2/2 checks passed
View alarms
us-east-1c
ec2-44-208-34-251.co...
EC-A
i-0b24410448adebb29
Running
t2.micro
2/2 checks passed
View alarms
us-east-1c
ec2-52-87-182-243.co...
i-055cafc9b00ced715
Running
m4.xlarge
2/2 checks passed
View alarms
us-east-1e
ec2-54-90-22-1.comput...
i-02f94c0fec253c513
Running
m4.xlarge
2/2 checks passed
View alarms
us-east-1e
ec2-35-153-33-29.com...
i-03632d9e7a0234d19
Running
m4.xlarge
2/2 checks passed
View alarms
us-east-1e
ec2-100-26-10-230.co...
i-00807dae3f8c55897
Running
m4.xlarge
2/2 checks passed
View alarms
us-east-1e
ec2-100-24-244-252.co...
i-09761729bd9532370
Running
m4.xlarge
2/2 checks passed
View alarms
us-east-1e
ec2-34-232-78-9.comp...
Select an instance
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

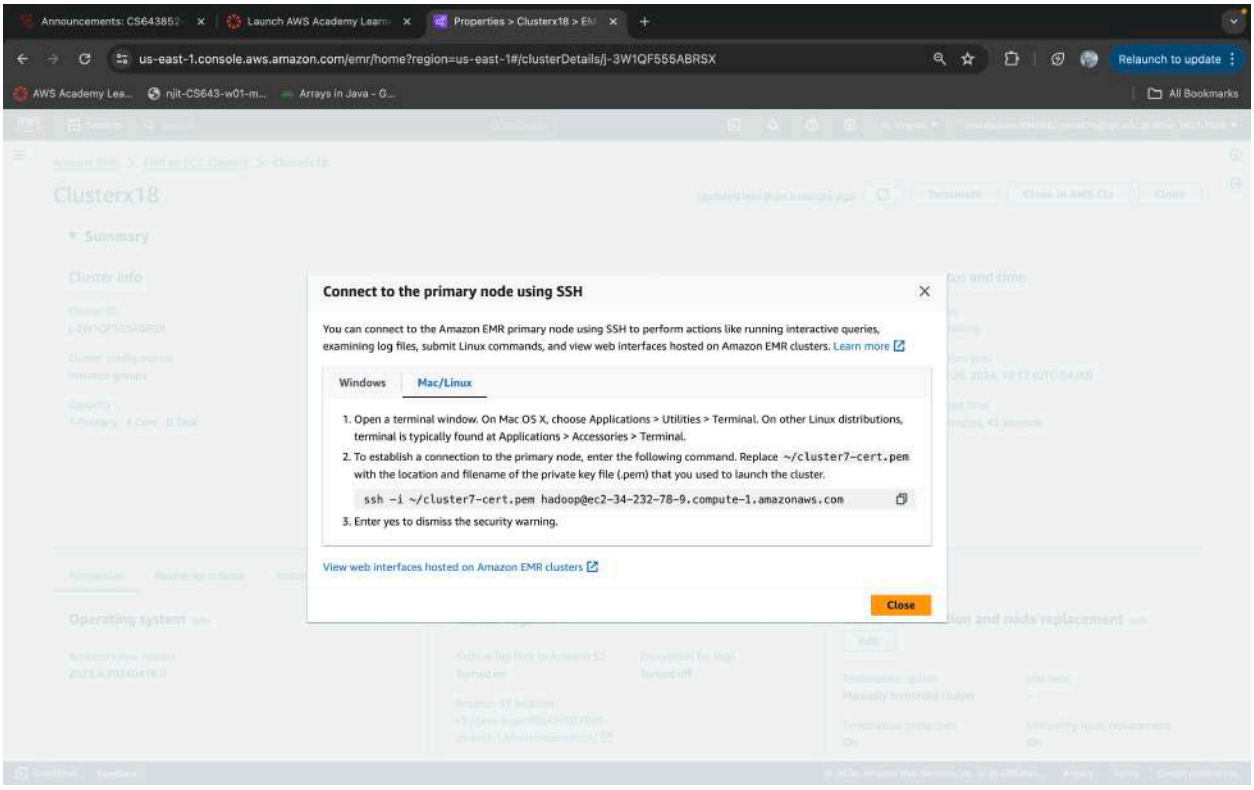
Create S3 bucket for storing input data csv files and trained ML Model:

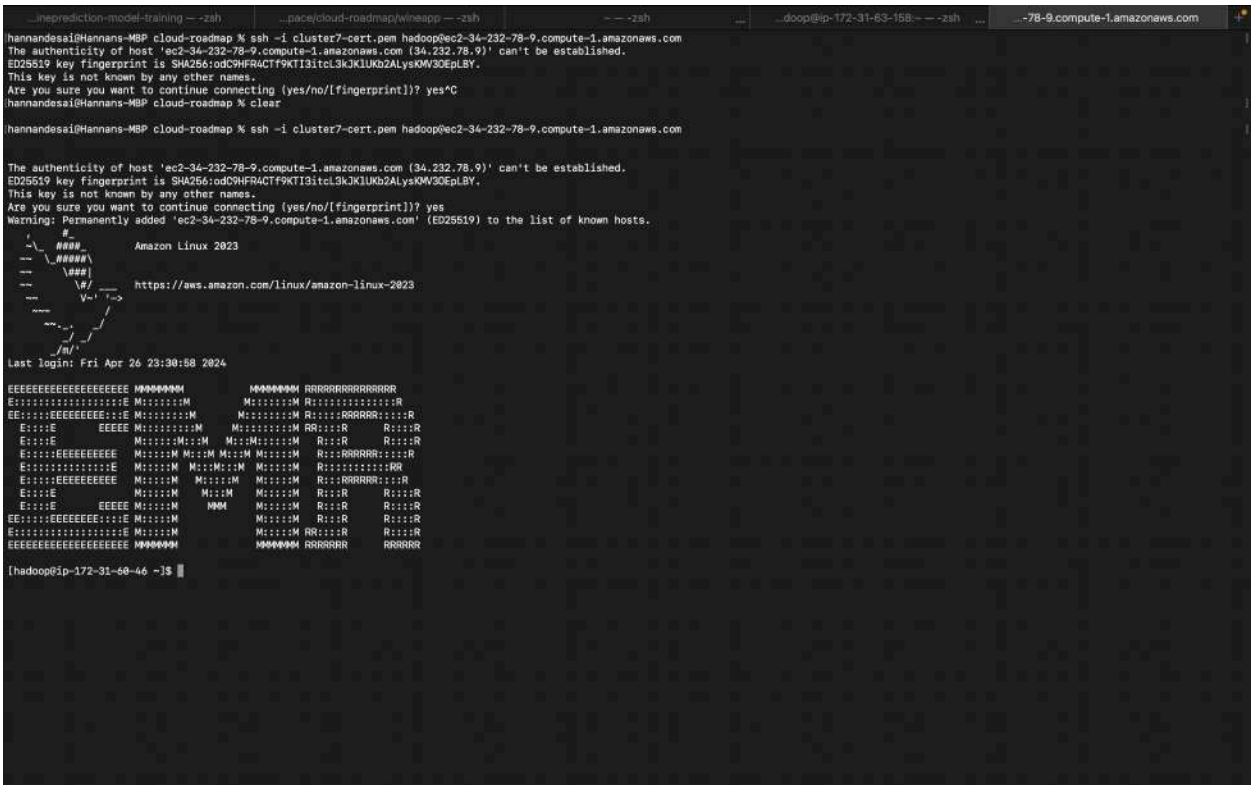


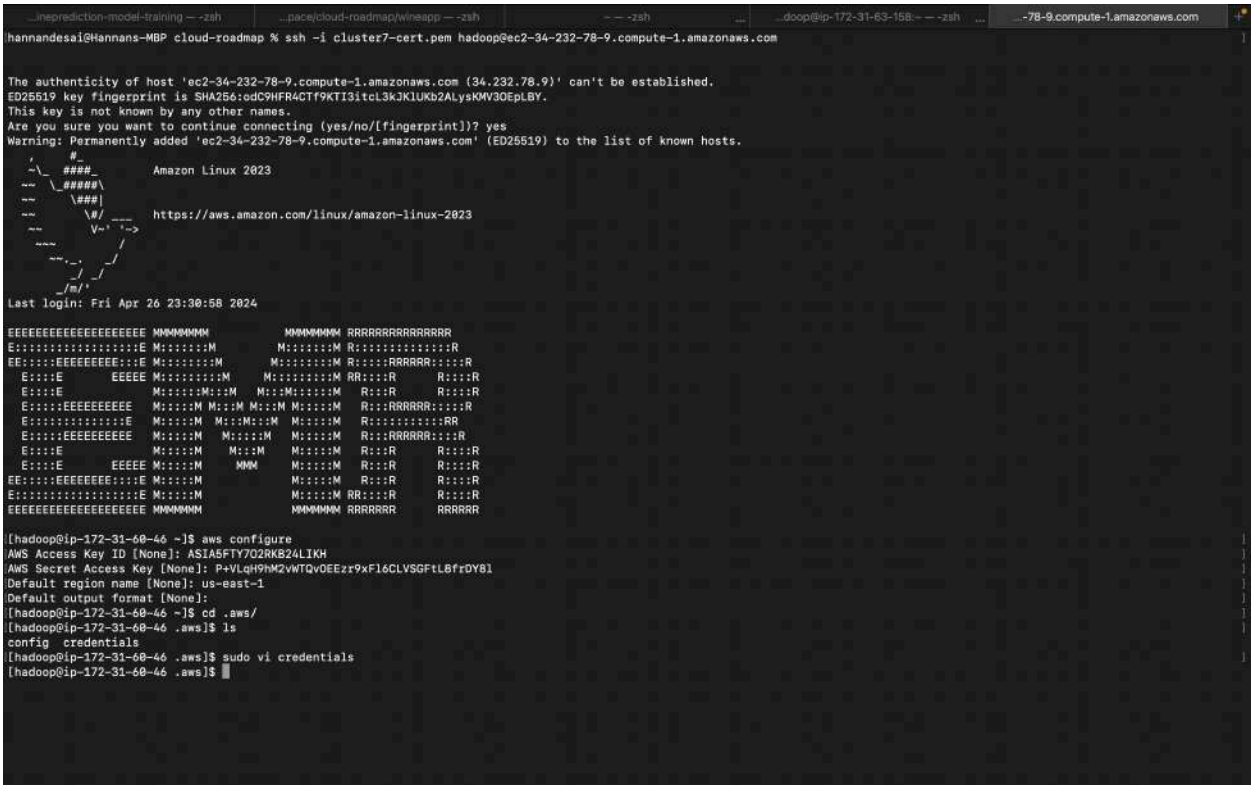
Run "wineprediction-model-training" application on created EMR Cluster:

1. **Connect to cluster:**


Announcements: CS643852
Launch AWS Academy Learn
Properties > Clusterx18 > EM
us-east-1.console.aws.amazon.com/emr/home?region=us-east-1#/clusterDetails/j-3W1QF555ABRSX
Relaunch to update
AWS Academy Lea...
njit-CS643-w01-m...
Arrays in Java - G...
All Bookmarks
Services
Search
[Option+S]
N. Virginia
vocabs/user3049882=mmd76@njit.edu @ 9054-1807-7858
Amazon EMR > EMR on EC2: Clusters > Clusterx18
Clusterx18
Updated less than a minute ago
Terminate
Clone in AWS CLI
Clone
Summary
Cluster info
Cluster ID
j-3W1QF555ABRSX
Cluster configuration
Instance groups
Capacity
1 Primary 4 Core 0 Task
Applications
Amazon EMR version
emr-7.1.0
Installed applications
Hadoop 3.3.6, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.8.0, Spark 3.5.0
Cluster management
Log destination in Amazon S3
aws-logs-905418077858-us-east-1/elasticmapreduce
Persistent application UIs
Spark History Server
YARN timeline server
Tez UI
Primary node public DNS
ec2-34-232-78-9.compute-1.amazonaws.com
Connect to the Primary node using SSH
Connect to the Primary node using SSM
Status and time
Status
Waiting
Creation time
April 26, 2024, 19:17 (UTC-04:00)
Elapsed time
25 minutes, 59 seconds
Properties
Bootstrap actions
Instances (Hardware)
Steps
Applications
Configurations
Monitoring
Events
Tags (0)
Operating system Info
Amazon Linux release
2023.4.20240416.0
Cluster logs Info
Archive log files to Amazon S3
Turned on
Amazon S3 location
s3://aws-logs-905418077858-us-east-1/elasticmapreduce/
Encryption for logs
Turned off
Cluster termination and node replacement Info
Edit
Termination option
Manually terminate cluster
Idle time
-
Termination protection
On
Unhealthy node replacement
On
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


Connect to the primary node using SSH
You can connect to the Amazon EMR primary node using SSH to perform actions like running interactive queries, examining log files, submit Linux commands, and view web interfaces hosted on Amazon EMR clusters. Learn more
Windows
Mac/Linux
1. Open a terminal window. On Mac OS X, choose Applications > Utilities > Terminal. On other Linux distributions, terminal is typically found at Applications > Accessories > Terminal.
2. To establish a connection to the primary node, enter the following command. Replace ~/cluster7-cert.pem with the location and filename of the private key file (.pem) that you used to launch the cluster.
ssh -i ~/cluster7-cert.pem hadoop@ec2-34-232-78-9.compute-1.amazonaws.com
3. Enter yes to dismiss the security warning.
View web interfaces hosted on Amazon EMR clusters
Close

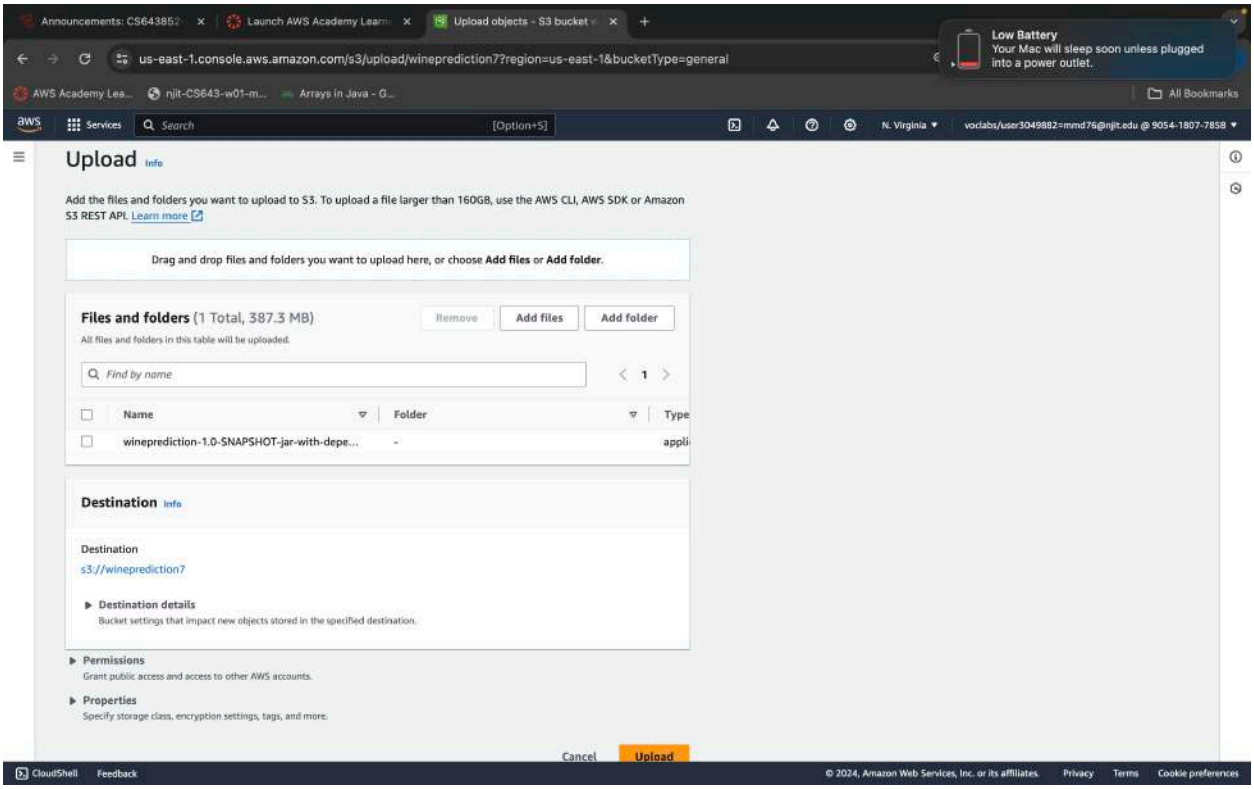

...ineprediction-model-training — -zsh
...pace/cloud-roadmap/wineapp — -zsh
~ — -zsh
...doop@ip-172-31-63-158:~ — -zsh
...-78-9.compute-1.amazonaws.com
hannandesai@Hannans-MBP cloud-roadmap % ssh -i cluster7-cert.pem hadoop@ec2-34-232-78-9.compute-1.amazonaws.com
The authenticity of host 'ec2-34-232-78-9.compute-1.amazonaws.com (34.232.78.9)' can't be established.
ED25519 key fingerprint is SHA256:odC9HFR4CTf9KTI3itcL3kJKlUKb2ALysKMV3OEpLBY.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes^C
hannandesai@Hannans-MBP cloud-roadmap % clear
hannandesai@Hannans-MBP cloud-roadmap % ssh -i cluster7-cert.pem hadoop@ec2-34-232-78-9.compute-1.amazonaws.com
The authenticity of host 'ec2-34-232-78-9.compute-1.amazonaws.com (34.232.78.9)' can't be established.
ED25519 key fingerprint is SHA256:odC9HFR4CTf9KTI3itcL3kJKlUKb2ALysKMV3OEpLBY.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-34-232-78-9.compute-1.amazonaws.com' (ED25519) to the list of known hosts.
Amazon Linux 2023
https://aws.amazon.com/linux/amazon-linux-2023
Last login: Fri Apr 26 23:30:58 2024
[hadoop@ip-172-31-60-46 ~]$

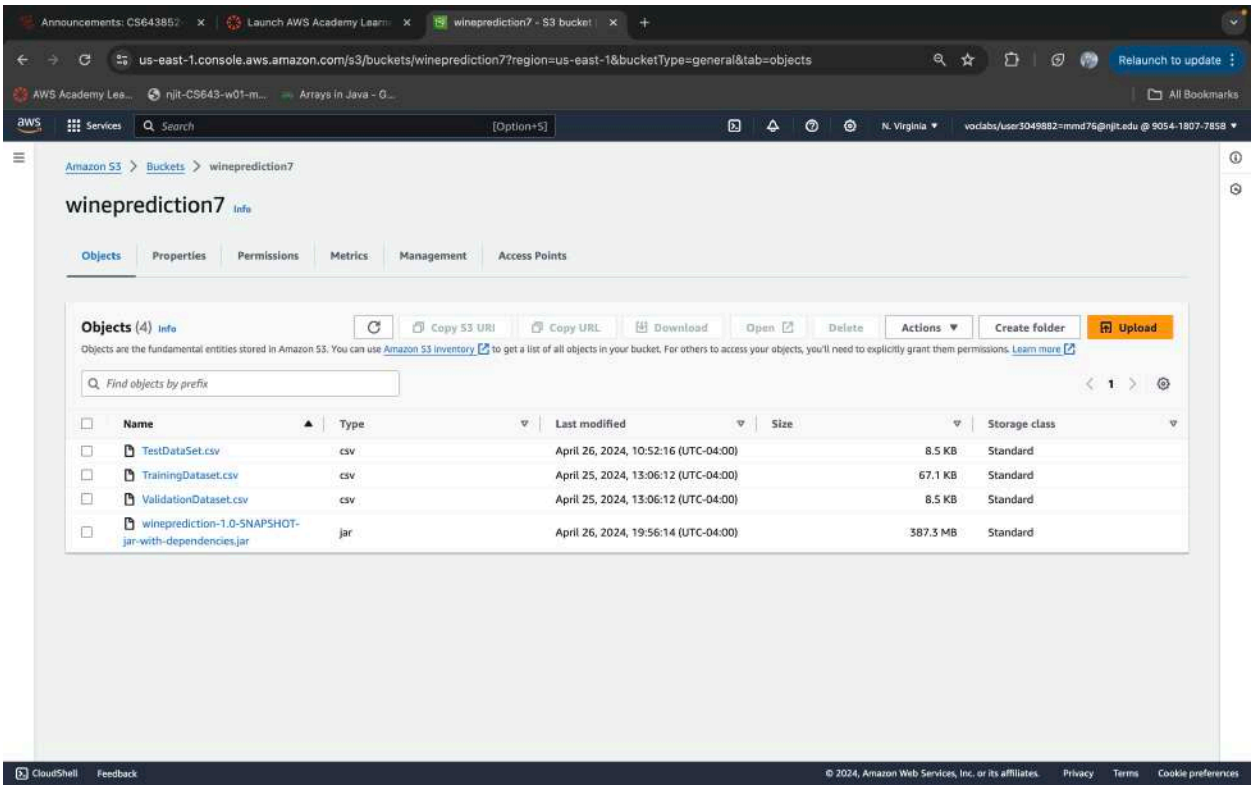

...ineprediction-model-training — -zsh
...pace/cloud-roadmap/wineapp — -zsh
~ — -zsh
...doop@ip-172-31-63-158:~ — -zsh
...-78-9.compute-1.amazonaws.com
hannandesai@Hannans-MBP cloud-roadmap % ssh -i cluster7-cert.pem hadoop@ec2-34-232-78-9.compute-1.amazonaws.com
The authenticity of host 'ec2-34-232-78-9.compute-1.amazonaws.com (34.232.78.9)' can't be established.
ED25519 key fingerprint is SHA256:odC9HFR4CTf9KTI3itcL3kJKlUKb2ALysKMV3OEpLBY.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-34-232-78-9.compute-1.amazonaws.com' (ED25519) to the list of known hosts.
Amazon Linux 2023
https://aws.amazon.com/linux/amazon-linux-2023
Last login: Fri Apr 26 23:30:58 2024
[hadoop@ip-172-31-60-46 ~]$ aws configure
AWS Access Key ID [None]: ASIA5FTY7O2RKB24LIKH
AWS Secret Access Key [None]: P+VLqH9hM2vWTQvOEEzr9xFl6CLVSGFtLBfrDY8l
Default region name [None]: us-east-1
Default output format [None]:
[hadoop@ip-172-31-60-46 ~]$ cd .aws/
[hadoop@ip-172-31-60-46 .aws]$ ls
config credentials
[hadoop@ip-172-31-60-46 .aws]$ sudo vi credentials
[hadoop@ip-172-31-60-46 .aws]$

2. **Create jar file of project:**

   Run “mvn clean package” in project directory, it will create jar file named “wineprediction-1.0-SNAPSHOT-jar-with-dependencies” in target folder.

   Upload this jar file on S3 bucket that we created in previous step:





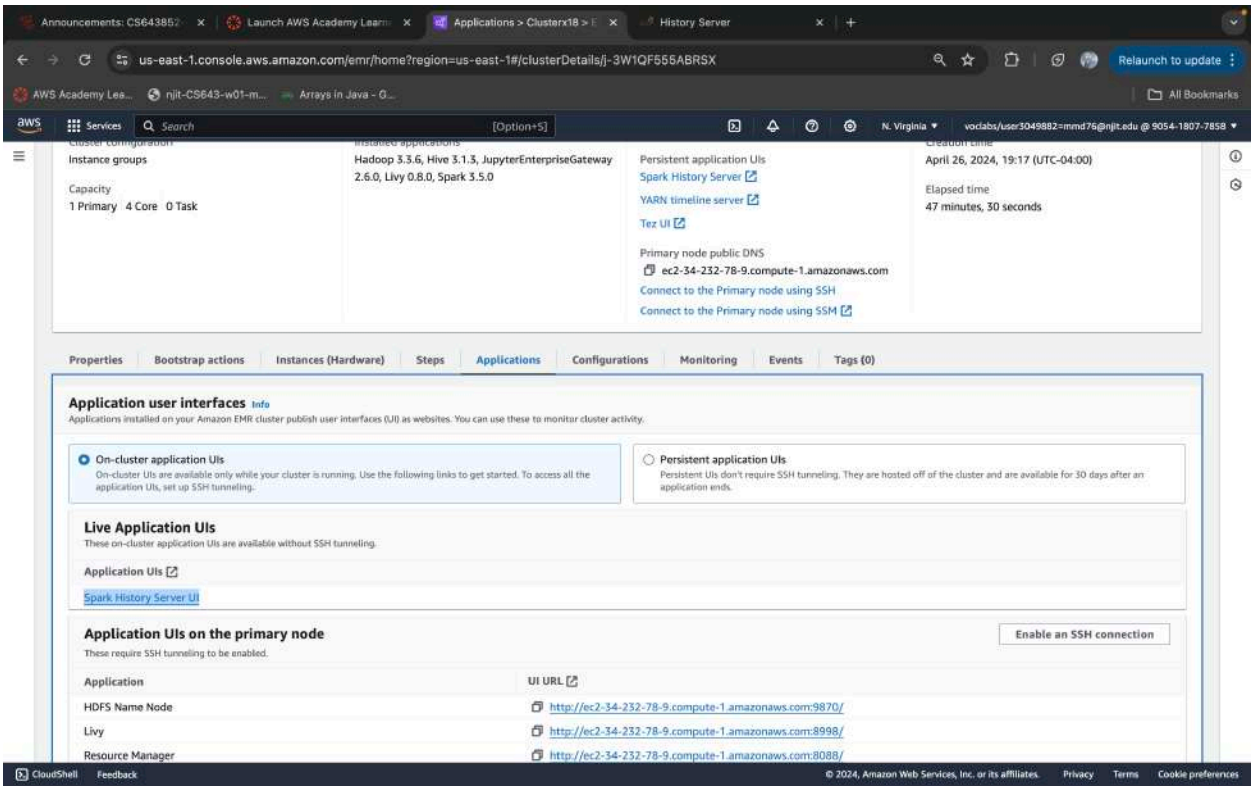
3. **Submit job to apche-spark cluster:**

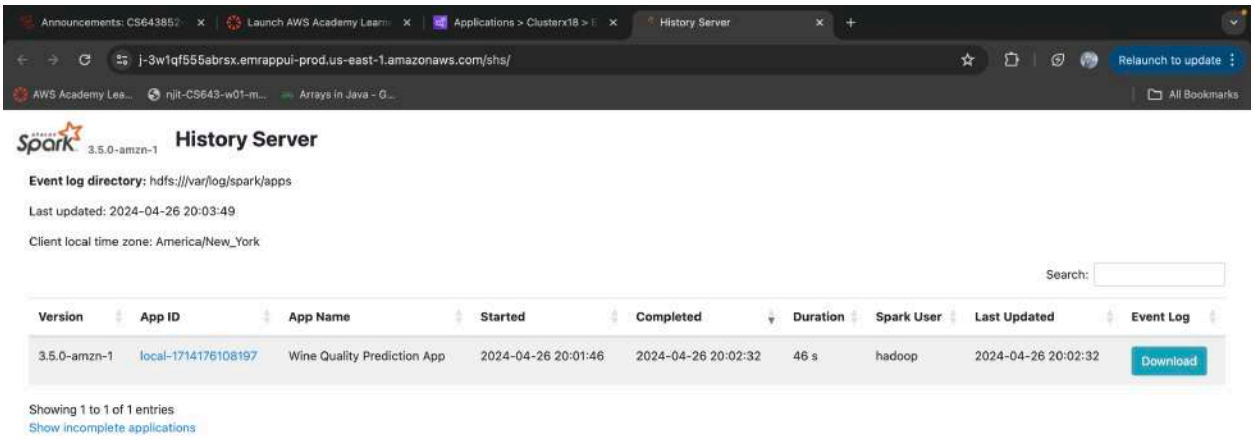
Run below command in terminal in which have connected our cluster:

spark-submit s3://wineprediction7/wineprediction-1.0-SNAPSHOT-jar-with-dependencies.jar

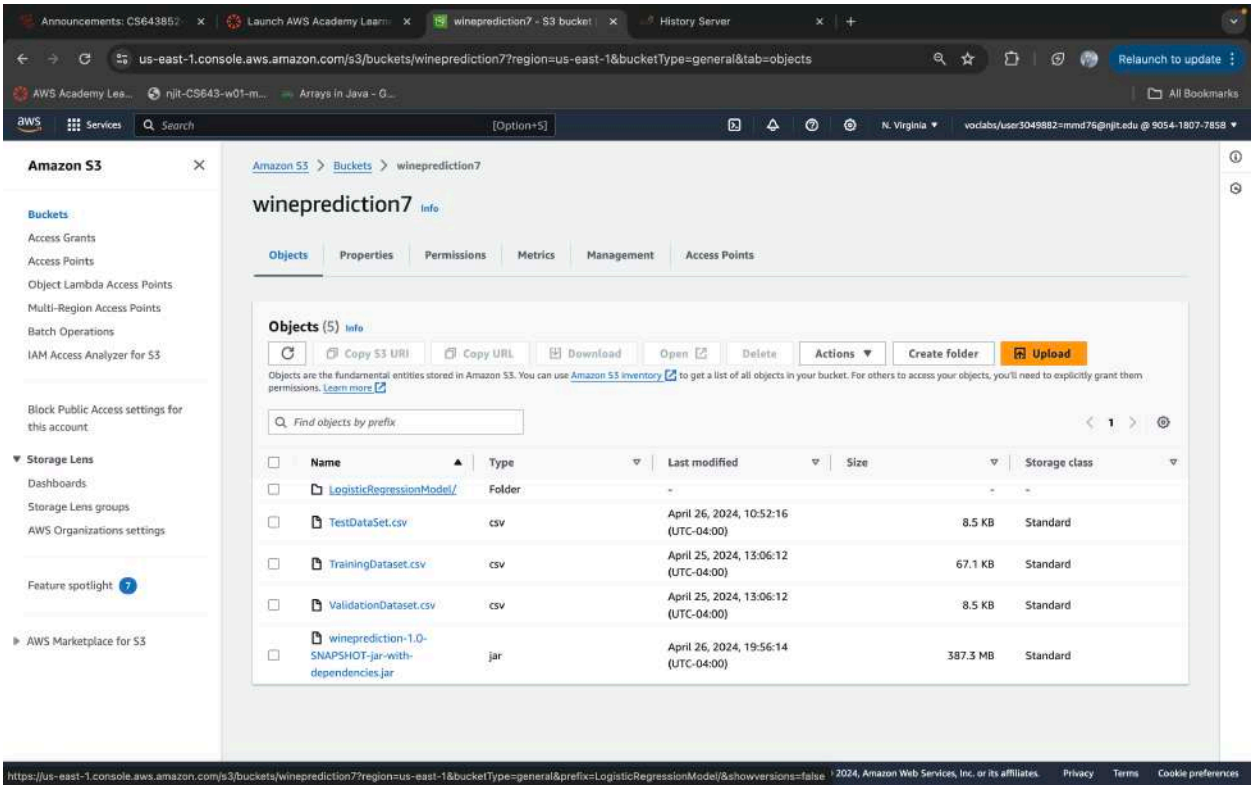
```
[hadoop@ip-172-31-60-46 ~]$ spark-submit s3://wineprediction7/wineprediction-1.0-SNAPSHOT-jar-with-dependencies.jar
Apr 27, 2024 12:01:30 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but /usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location

24/04/27 00:01:46 INFO SparkContext: Running Spark version 3.5.0-amzn-1
24/04/27 00:01:46 INFO SparkContext: OS info Linux, 6.1.84-99.169.amzn2023.x86_64, amd64
24/04/27 00:01:46 INFO SparkContext: Java version 17.0.10
24/04/27 00:01:46 INFO ResourceUtils: ==============================================================
24/04/27 00:01:46 INFO ResourceUtils: No custom resources configured for spark.driver.
24/04/27 00:01:46 INFO ResourceUtils: ==============================================================
24/04/27 00:01:46 INFO SparkContext: Submitted application: Wine Quality Prediction App
24/04/27 00:01:46 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 4, script: , vendor: , memory -> name: memory, amount: 3072, script: , vendor
: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
24/04/27 00:01:46 INFO ResourceProfile: Limiting resource is cpus at 4 tasks per executor
24/04/27 00:01:47 INFO ResourceProfileManager: Added ResourceProfile id: 0
24/04/27 00:01:47 INFO SecurityManager: Changing view acls to: hadoop
24/04/27 00:01:47 INFO SecurityManager: Changing modify acls to: hadoop
24/04/27 00:01:47 INFO SecurityManager: Changing view acls groups to:
24/04/27 00:01:47 INFO SecurityManager: Changing modify acls groups to:
24/04/27 00:01:47 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: hadoop; groups with view permissions: EMPTY; users with modify permissions:
 hadoop; groups with modify permissions: EMPTY
24/04/27 00:01:47 INFO Utils: Successfully started service 'sparkDriver' on port 38711.
24/04/27 00:01:47 INFO SparkEnv: Registering MapOutputTracker
24/04/27 00:01:47 INFO SparkEnv: Registering BlockManagerMaster
24/04/27 00:01:47 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
24/04/27 00:01:47 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
24/04/27 00:01:47 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
24/04/27 00:01:47 INFO DiskBlockManager: Created local directory at /mnt/tmp/blockmgr-32d71d73-dabb-4e99-a226-dd380df901de
24/04/27 00:01:47 INFO MemoryStore: MemoryStore started with capacity 1048.8 MiB
24/04/27 00:01:47 INFO SparkEnv: Registering OutputCommitCoordinator
24/04/27 00:01:47 INFO SubResultCacheManager: Sub-result caches are disabled.
24/04/27 00:01:48 INFO JettyUtils: Start Jetty 0.0.0.0:4040 for SparkUI
24/04/27 00:01:48 INFO Utils: Successfully started service 'SparkUI' on port 4040.
24/04/27 00:01:48 INFO SparkContext: Added JAR s3://wineprediction7/wineprediction-1.0-SNAPSHOT-jar-with-dependencies.jar at s3://wineprediction7/wineprediction-1.0-SNAPSHOT-jar-with-dependencies.jar with
 timestamp 1714176106904
24/04/27 00:01:48 INFO Executor: Starting executor ID driver on host ip-172-31-60-46.ec2.internal
24/04/27 00:01:48 INFO Executor: OS info Linux, 6.1.84-99.169.amzn2023.x86_64, amd64
24/04/27 00:01:48 INFO Executor: Java version 17.0.10
24/04/27 00:01:48 INFO Executor: Starting executor with user classpath (userClassPathFirst = false): 'file:/usr/lib/hadoop-lzo/lib/*,file:/usr/lib/hadoop/hadoop-aws.jar,file:/usr/share/aws/aws-java-sdk/*,
file:/usr/share/aws/aws-java-sdk-v2/*,file:/usr/share/aws/emr/goodies/lib/emr-spark-goodies.jar,file:/usr/share/aws/emr/security/conf,file:/usr/share/aws/emr/security/lib/*,file:/usr/share/aws/redshift/jd
bc/*,file:/usr/share/aws/redshift/spark-redshift/lib/*,file:/usr/share/aws/kinesis/spark-sql-kinesis/lib/*,file:/usr/share/aws/hmclient/lib/aws-glue-datacatalog-spark-client.jar,file:/usr/share/java/Hive-
JSON-Serde/hive-openx-serde.jar,file:/usr/share/aws/emr/s3select/lib/emr-s3-select-spark-connector.jar,file:/docker/usr/lib/hadoop-lzo/lib/*,file:/docker/usr/lib/hadoop/hadoop-aws.jar,file:/docker/usr/sha
re/aws/aws-java-sdk/*,file:/docker/usr/share/aws/aws-java-sdk-v2/*,file:/docker/usr/share/aws/emr/goodies/lib/emr-spark-goodies.jar,file:/docker/usr/share/aws/emr/security/conf,file:/docker/usr/share/aws/
emr/security/lib/*,file:/docker/usr/share/aws/redshift/jdbc/*,file:/docker/usr/share/aws/redshift/spark-redshift/lib/*,file:/docker/usr/share/aws/kinesis/spark-sql-kinesis/lib/*,file:/docker/usr/share/aws
/hmclient/lib/aws-glue-datacatalog-spark-client.jar,file:/docker/usr/share/java/Hive-JSON-Serde/hive-openx-serde.jar,file:/docker/usr/share/aws/emr/s3select/lib/emr-s3-select-spark-connector.jar,file:/hom
e/hadoop/conf,file:/home/hadoop/emr-spark-goodies.jar,file:/home/hadoop/*,file:/home/hadoop/aws-glue-datacatalog-spark-client.jar,file:/home/hadoop/hive-openx-serde.jar,file:/home/hadoop/hadoop-aws.jar,fi
le:/home/hadoop/emr-s3-select-spark-connector.jar'
24/04/27 00:01:48 INFO Executor: Created or updated repl class loader org.apache.spark.util.MutableURLClassLoader@6124dda2 for default.
24/04/27 00:01:48 INFO Executor: Fetching s3://wineprediction7/wineprediction-1.0-SNAPSHOT-jar-with-dependencies.jar with timestamp 1714176106904
24/04/27 00:01:48 INFO S3NativeFileSystem: Opening 's3://wineprediction7/wineprediction-1.0-SNAPSHOT-jar-with-dependencies.jar' for reading
24/04/27 00:01:48 INFO Utils: Fetching s3://wineprediction7/wineprediction-1.0-SNAPSHOT-jar-with-dependencies.jar to /mnt/tmp/spark-87e5385c-c5ad-4863-b83e-1f8dfd4484f4/userFiles-f9bfb911-3ad3-463c-b9ab-2
85becdecffc/fetchFileTemp14513335490333433517.tmp
24/04/27 00:01:52 INFO Executor: Adding file:/mnt/tmp/spark-87e5385c-c5ad-4863-b83e-1f8dfd4484f4/userFiles-f9bfb911-3ad3-463c-b9ab-285becdecffc/wineprediction-1.0-SNAPSHOT-jar-with-dependencies.jar to cla
ss loader default
24/04/27 00:01:52 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 37847.
24/04/27 00:01:52 INFO NettyBlockTransferService: Server created on ip-172-31-60-46.ec2.internal:37847
24/04/27 00:01:52 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
```

We can see out submitted job using “Spark Histroy Server UI” link:



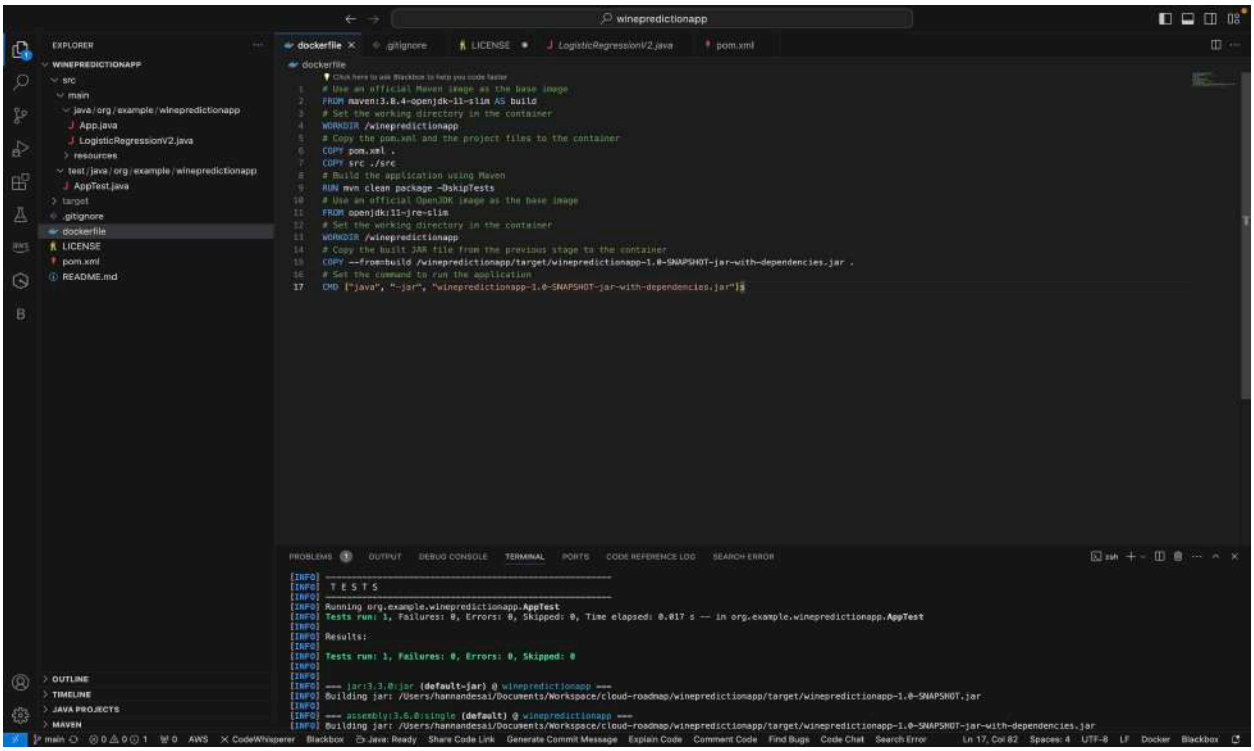


So once this job is finished it will save the Trained model named “LinerRegressionModel” in S3 bucket:

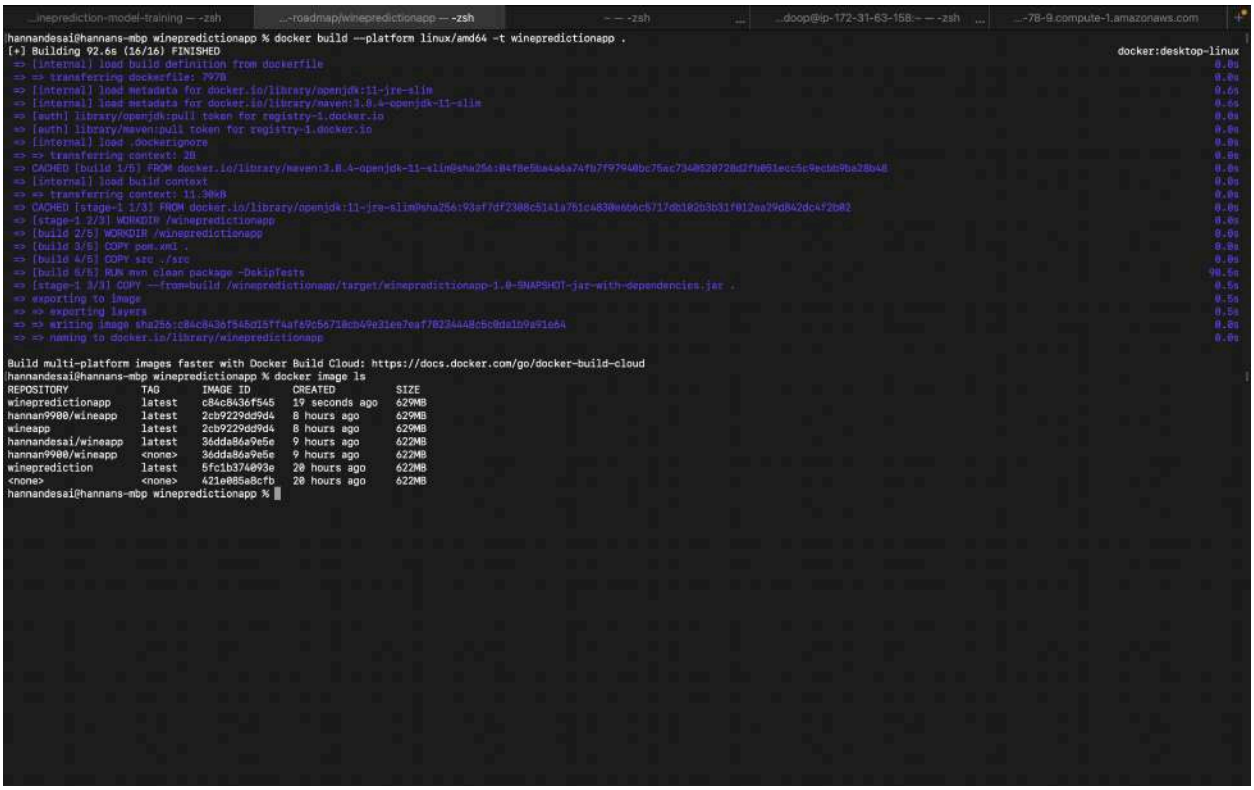


Create DOCKER image for prediction app:

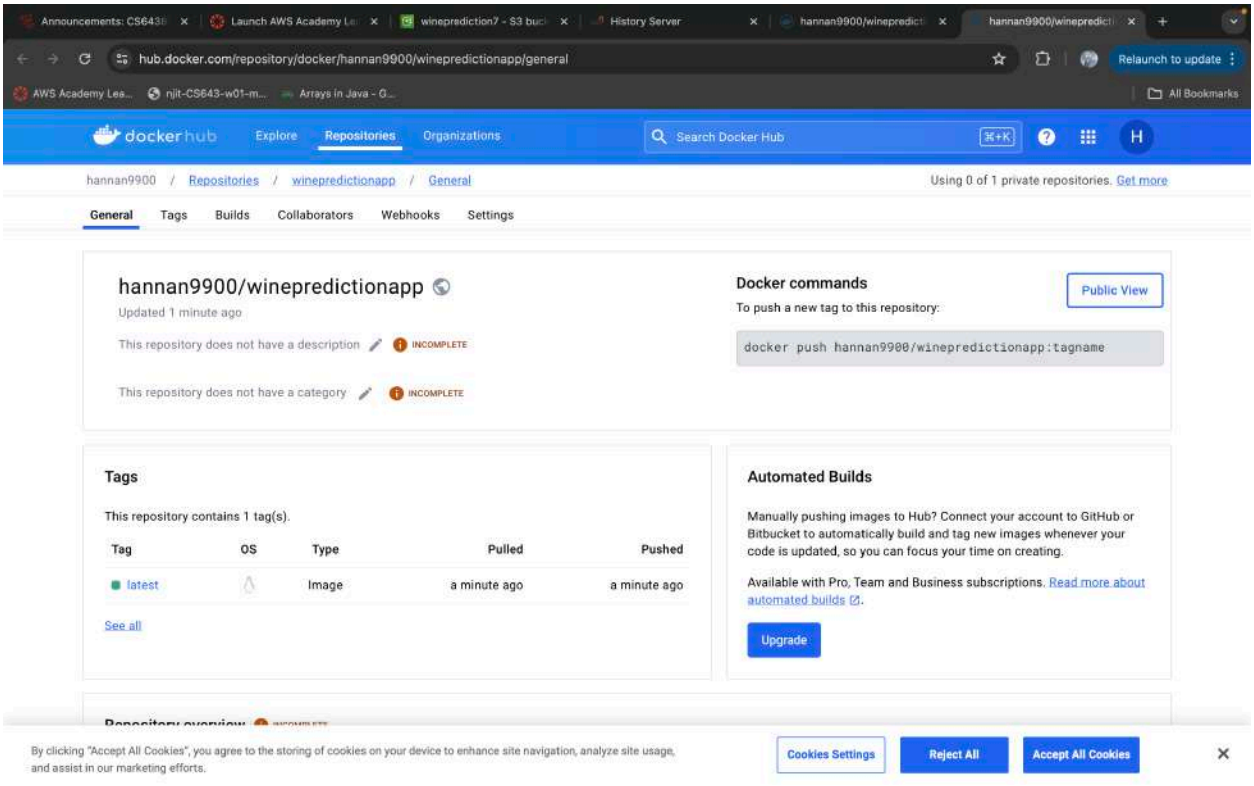
1. **Create docker file in winepredictionapp root folder:**



2. **Build the docker image:**



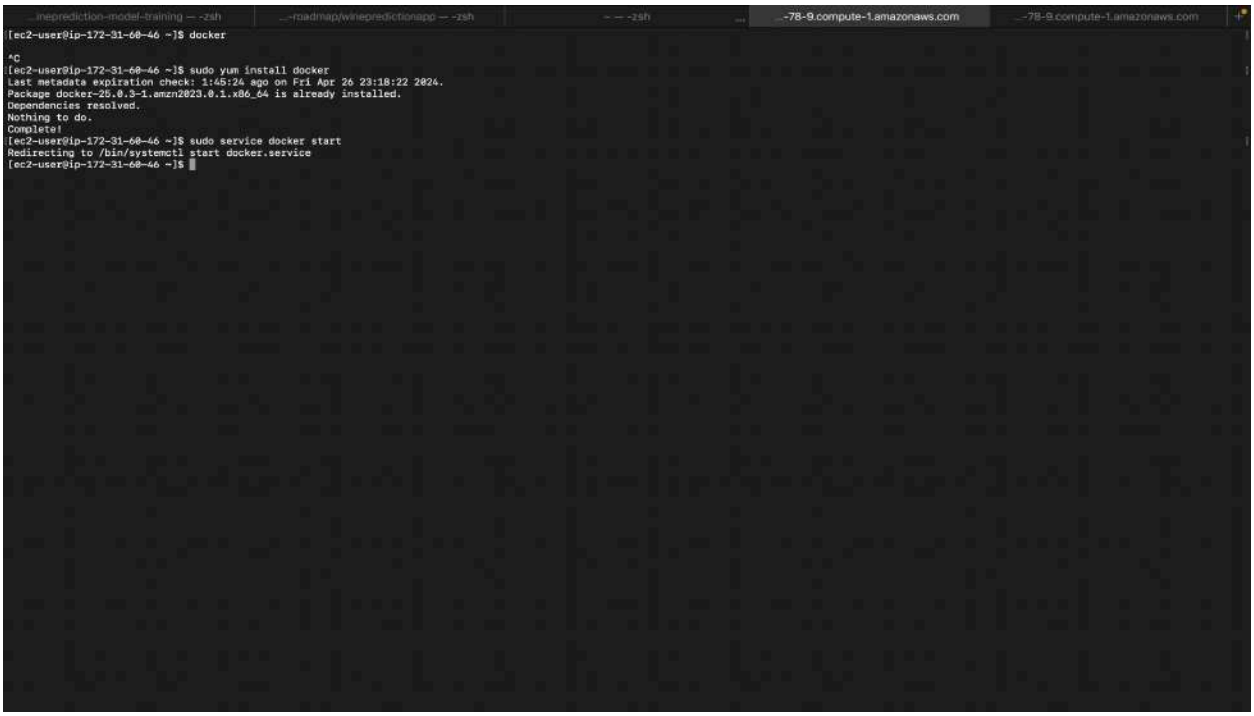
3. Push to created image to public docker hub:



Run this docker image in AWS EC2 instance:

1. **Connect to an EC2 instance:**

2. **Install docker on instance:**



3. Fetch docker image from docker hub to instance:

4. Run Docker image:

To run this prediction app without docker:

Run following command in project root directory:

mvn clean package

```
mvn exec:java –
Dexec.mainClass="org.example.winepredictionapp.App";
```