

## **DTSC 701: INTRODUCTION TO BIG DATA**

### **PROJECT ASSIGNMENT**

Names:

**AISHWARYA KANAKAMEDALA (ID: 1322532)**

**TOYAZ VAMSI INAGANTI (ID: 1316806)**

1. A data lake can include structured data from relational databases (rows and columns), semi-structured data (CSV, logs, XML, JSON), unstructured data (emails, documents, PDFs) and binary data (images, audio, video). Therefore a CSV as data lake downloaded from Kaggle, which has fuel consumption from year 2000 to 2022, with more than 27000 rows.

1	ticker	commodity	date	open	high	low	close	volume
2	CL=F	Crude Oil	8/23/2000	31.95000076	32.79999924	31.95000076	32.04999924	79385
3	CL=F	Crude Oil	8/24/2000	31.89999962	32.24000168	31.39999962	31.62999916	72978
4	CL=F	Crude Oil	8/25/2000	31.70000076	32.09999847	31.31999969	32.04999924	44601
5	CL=F	Crude Oil	8/28/2000	32.04000092	32.91999817	31.86000061	32.86999893	46770
6	CL=F	Crude Oil	8/29/2000	32.81999969	33.02999878	32.56000137	32.72000122	49131
7	CL=F	Crude Oil	8/30/2000	32.75	33.40000153	32.09999847	33.40000153	79214
8	CL=F	Crude Oil	8/31/2000	33.25	33.70000076	32.97000122	33.09999847	56895
9	CL=F	Crude Oil	9/1/2000	33.04999924	33.45000076	32.75	33.38000107	45869
10	CL=F	Crude Oil	9/5/2000	33.95000076	33.99000168	33.41999817	33.79999924	55722
11	CL=F	Crude Oil	9/6/2000	33.99000168	34.95000076	33.83000183	34.95000076	74692
12	CL=F	Crude Oil	9/7/2000	34.5	35.5	34.45000076	35.33000183	74105
13	CL=F	Crude Oil	9/8/2000	34.54999924	34.77999878	33.40000153	33.70000076	88415
14	CL=F	Crude Oil	9/11/2000	33.79999924	35.84999847	33.75	35.09999847	101518
15	CL=F	Crude Oil	9/12/2000	35.45000076	35.5	34.09999847	34.20000076	91911
16	CL=F	Crude Oil	9/13/2000	34	34.74000168	33.5	33.79999924	94630
17	CL=F	Crude Oil	9/14/2000	33.77999878	34.5	33.11999893	34.09999847	98068
18	CL=F	Crude Oil	9/15/2000	34.5	36.09999847	34.45000076	35.84999847	85839
19	CL=F	Crude Oil	9/18/2000	36.20000076	37.15000153	36.15000153	36.88000107	59663
20	CL=F	Crude Oil	9/19/2000	36.54999924	37	36.15000153	36.5	62731
21	CL=F	Crude Oil	9/20/2000	37.5	37.79999924	36.5	37.5	119080
22	CL=F	Crude Oil	9/21/2000	34.65000153	35.5	33.34999847	33.95000076	110851
23	CL=F	Crude Oil	9/22/2000	34	34.40000153	32.5	32.65000153	85083

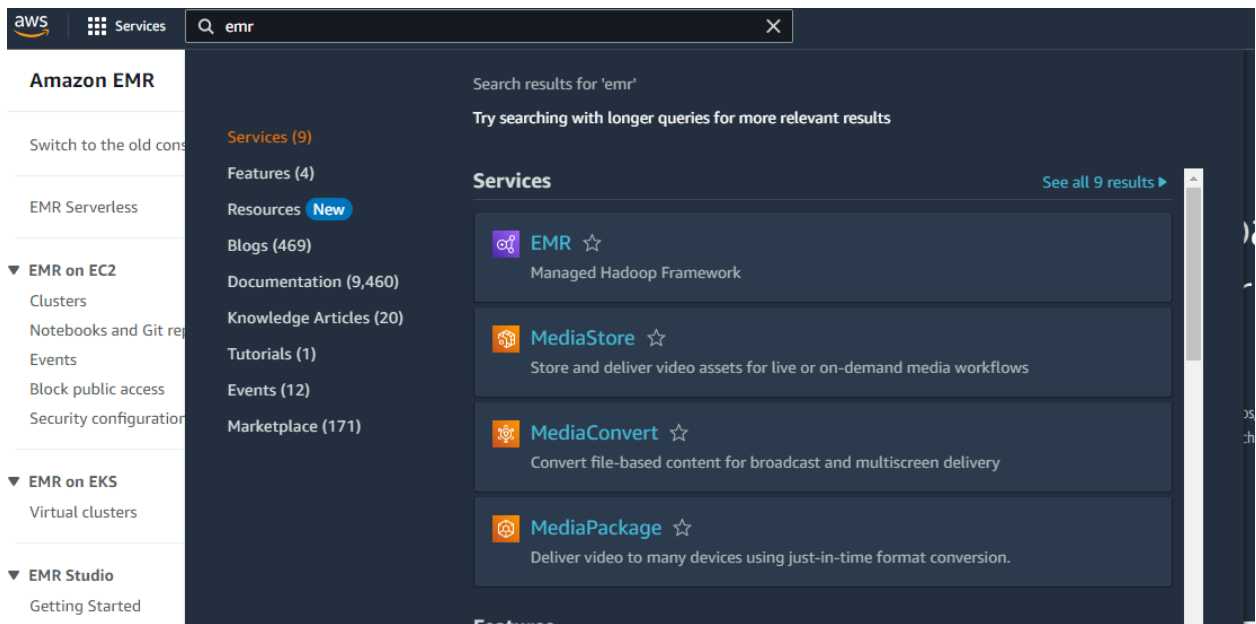
2. Register or sign in an amazon web services (AWS).

The screenshot shows the Amazon EMR console interface. On the left is a navigation sidebar with the following sections:

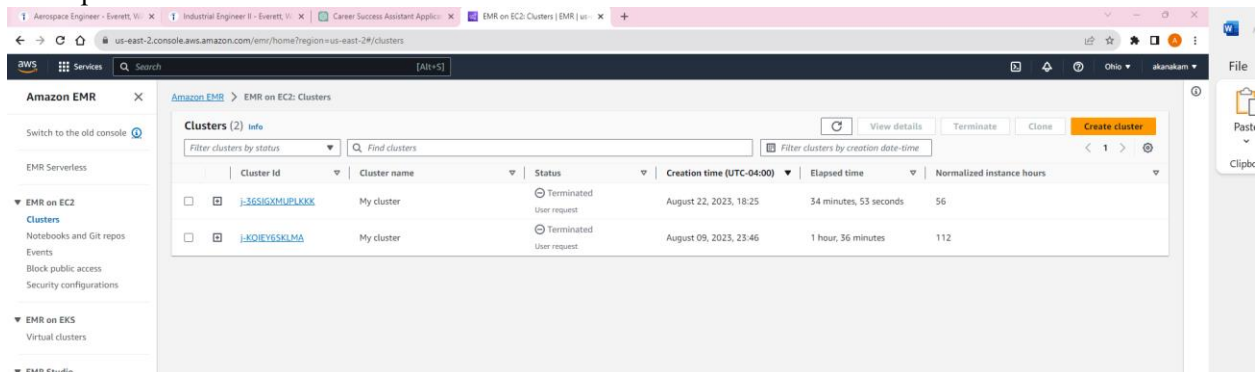
- Amazon EMR (with a close button)
- Switch to the old console (with an information icon)
- EMR Serverless
- EMR on EC2
  - Clusters
  - Notebooks and Git repos
  - Events
  - Block public access
  - Security configurations
- EMR on EKS
  - Virtual clusters
- EMR Studio
  - Getting Started
  - Studios
  - Workspaces (Notebooks)
- What's New
- Compact mode (toggle switch)

The main content area features a large header for 'Amazon EMR' with the tagline 'Easily run and scale Apache Spark, Apache Hive, Presto, and other big data workloads.' Below this is a brief description of Amazon EMR as a cloud big data platform. At the bottom, there is a 'How it works' section containing a video player titled 'An introduction to Amazon EMR - Amazon Web Services' with a 'Copy link' button.

3. Navigate to EMR in the search bar to create a new cluster



4. set up a Spark cluster on AWS using Amazon EMR (Elastic MapReduce) or by deploying your own Spark cluster on EC2 instances.



5. Name a new cluster, where we choose Spark or customize depending on the requirement as Application bundle, and cluster configuration as Primary(m5.xlarge), Core (m5.xlarge), Task (m5.xlarge) and increasing Cluster scaling and provisioning option as Core size as 3 instances and Task size as 3 instances. For creating a cluster for machine learning purpose, we select additional resources in the customized bundle as shown in fig below

Create cluster [Info](#)

## Name and applications [Info](#)

Name

Myproject

Amazon EMR release [Info](#)

A release contains a set of applications which can be installed on your cluster.

emr-6.12.0

Application bundle

Spark

Core Hadoop

Flink

HBase

Presto

Trino

Custom

Applications included in bundle

Spark 3.4.0 on Hadoop 3.3.3 YARN with and Zeppelin 0.10.1

AWS Glue Data Catalog settings

Use the AWS Glue Data Catalog to provide an external metastore for your application.

☐ Use for Spark table metadata

Operating system options [Info](#)

☒ Amazon Linux release

☐ Custom Amazon Machine Image (AMI)

☒ Automatically apply latest Amazon Linux updates

## Cluster configuration [info](#)

Choose a configuration method for the primary, core, and task node groups for your cluster.

☒ **Instance groups**  
Choose an instance type for each group.

☐ **Instance fleets**  
Choose an availability of instance types within each group.

## Summary [Info](#)

# Name and applications

Name

Myproject

Amazon EMR release

emr-6.12.0

Application bundle

Spark

## Cluster configuration

Instance groups


Primary (m5.xlarge), Core (m5.xlarge), Task (m5.xlarge)

## Cluster scaling and provisioning option

Provisioning configuration

Core size: 3 instances

Task size: 3 instances

 **Configure IAM roles**  
You must choose a service role and instance profile before you create this cluster.








Choose IAM roles

Cancel

Create cluster

Or

## Application bundle

Spark	Core Hadoop	Flink	HBase	Presto	Trino	Custom
						

▼ Customize your application bundle

Applications included in bundle

- |                                                 |                                                                    |                                                       |
|-------------------------------------------------|--------------------------------------------------------------------|-------------------------------------------------------|
| <input type="checkbox"/> Flink 1.17.0           | <input type="checkbox"/> Ganglia 3.7.2                             | <input type="checkbox"/> HBase 2.4.17                 |
| <input type="checkbox"/> HCatalog 3.1.3         | <input checked="" type="checkbox"/> Hadoop 3.3.3                   | <input checked="" type="checkbox"/> Hive 3.1.3        |
| <input checked="" type="checkbox"/> Hue 4.11.0  | <input checked="" type="checkbox"/> JupyterEnterpriseGateway 2.6.0 | <input checked="" type="checkbox"/> JupyterHub 1.4.1  |
| <input type="checkbox"/> Livy 0.7.1             | <input type="checkbox"/> MXNet 1.9.1                               | <input type="checkbox"/> Oozie 5.2.1                  |
| <input type="checkbox"/> Phoenix 5.1.3          | <input checked="" type="checkbox"/> Pig 0.17.0                     | <input type="checkbox"/> Presto 0.281                 |
| <input checked="" type="checkbox"/> Spark 3.4.0 | <input type="checkbox"/> Sqoop 1.4.7                               | <input checked="" type="checkbox"/> TensorFlow 2.11.0 |
| <input type="checkbox"/> Tez 0.10.2             | <input type="checkbox"/> Trino 414                                 | <input type="checkbox"/> Zeppelin 0.10.1              |
| <input type="checkbox"/> ZooKeeper 3.5.10       |                                                                    |                                                       |

### 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

Operating system options [Info](#)

- ☒ Amazon Linux release
- ☐ Custom Amazon Machine Image (AMI)
- ☒ Automatically apply latest Amazon Linux updates

6. Scale the cluster to the size of 3 instances for core and task nodes.

**Cluster scaling and provisioning option** [Info](#)  
Amazon EMR console only supports EMR-managed scaling. To create a cluster with auto-scaling, use CLI or SDK.

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.

**Provisioning configuration**  
Set the size of your core and task instance groups. Amazon EMR attempts to provision this capacity when you launch your cluster.

Name	Instance type	Instance(s) size	Use Spot purchasing option
Core	m5.xlarge	<input type="text" value="3"/>	<input type="checkbox"/>
Task - 1	m5.xlarge	<input type="text" value="3"/>	<input type="checkbox"/>

7. Create or security configuration and EC2 key pair as well as Identity and Access Management (IAM) roles

**Security configuration and EC2 key pair - optional** [Info](#)

**Security configuration**  
Select your cluster encryption, authentication, authorization, and instance metadata service settings.

**Amazon EC2 key pair for SSH to the cluster** [Info](#)

**⚠** You haven't entered an EC2 key. If you're outside a VPN and want to enable SSH or use Hue SQL assistant with this cluster, you must enter an EC2 key.

**Identity and Access Management (IAM) roles** [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**

**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.

8. Create a key pair by clicking on **create key pair** as shown in the picture below and then browse it and select it for the EC2 pair. This step is completely option

EC2 > Key pairs > Create key pair

## Create key pair [Info](#)

**Key pair**  
A key pair, consisting of a private key and a public key, is a set of security credentials that you use to prove your identity when connecting to an instance.

Name  
  
The name can include up to 255 ASCII characters. It can't include leading or trailing spaces.

Key pair type [Info](#)  
☒ RSA ☐ ED25519

Private key file format  
☐ .pem  
For use with OpenSSH  
☒ .ppk  
For use with PuTTY

Tags - *optional*  
No tags associated with the resource.  
[Add new tag](#)  
You can add up to 50 more tags.

[Cancel](#) [Create key pair](#)

9. Next step is to Create a service role or use an existing service role. Below are the pictures showing how to create a VPC and subnet and security group for the first time. By creating new **Virtual Private Cloud (VPC)** and **subnet** and **Security group**. Search VPC on search bar and get navigated to the picture below

**VPC dashboard** [EC2 Global View](#) [New](#)

Filter by VPC:  
[Select a VPC](#)

**Virtual private cloud**

- Your VPCs [New](#)
- Subnets
- Route tables
- Internet gateways
- Egress-only internet gateways
- DHCP option sets
- Elastic IPs
- Managed prefix lists

**Create VPC** [Launch EC2 Instances](#)

Note: Your Instances will launch in the US East region.

### Resources by Region [Refresh Resources](#)

You are using the following Amazon VPC resources

Resource	US East 1	US East 0
VPCs	1	0
Subnets	3	0
Route Tables	1	0
Internet Gateways	1	0
Egress-only Internet	0	0
NAT Gateways	0	0
VPC Peering Connections	0	0
Network ACLs	1	0
Security Groups	4	0

**Service Health**  
[View complete service health details](#)

**Settings**  
[Zones](#)  
[Console Experiments](#)

**Additional Information**  
[VPC Documentation](#)  
[All VPC Resources](#)  
[Forums](#)  
[Report an Issue](#)

**AWS Network Manager**  
AWS Network Manager provides tools and features to help you manage and monitor your network on AWS. Network Manager makes it easier to perform connectivity management, network

Choose the option of VPC and more, and name the VPC. The option of VPC and more will create the subnet and security group key also as shown in the below pictures.

[VPC](#) > [Your VPCs](#) > Create VPC

## Create VPC [Info](#)

A VPC is an isolated portion of the AWS Cloud populated by AWS objects, such as Amazon EC2 instances.

### VPC settings

Resources to create [Info](#)  
Create only the VPC resource or the VPC and other networking resources.

☒ VPC only ☐ VPC and more

Name tag - *optional*  
Creates a tag with a key of 'Name' and a value that you specify.

IPv4 CIDR block [Info](#)

☒ IPv4 CIDR manual input ☐ IPAM-allocated IPv4 CIDR block

IPv4 CIDR

[VPC](#) > [Your VPCs](#) > [Create VPC](#) > Create VPC resources

## Create VPC workflow

✔ Success

▼ Details

- ✔ Create VPC: [vpc-09dcb33e1bb2cc7ce](#)
- ✔ Enable DNS hostnames
- ✔ Enable DNS resolution
- ✔ Verifying VPC creation: [vpc-09dcb33e1bb2cc7ce](#)
- ✔ Create S3 endpoint: [vpce-00ad5da4630f75bd8](#)
- ✔ Create subnet: [subnet-04f67a305c004e76c](#)
- ✔ Create subnet: [subnet-0f01f7cc66f82f588](#)
- ✔ Create subnet: [subnet-00f85f08cbc5487e5](#)
- ✔ Create subnet: [subnet-0790e6848682d678e](#)
- ✔ Create internet gateway: [igw-0103b4ad5fa2fa72e](#)
- ✔ Attach internet gateway to the VPC
- ✔ Create route table: [rtb-08126931f8ab64526](#)
- ✔ Create route

The image below shows the VPC details that is created and can be used in the cluster

VPC > Your VPCs > vpc-09dcb33e1bb2cc7ce

## vpc-09dcb33e1bb2cc7ce / project-vpc

Actions ▼

Details Info

VPC ID vpc-09dcb33e1bb2cc7ce	State Available	DNS hostnames Enabled	DNS resolution Enabled
Tenancy Default	DHCP option set dopt-04f9ba1605b89bc1a	Main route table rtb-0c74143cab02c6c1c	Main network ACL acl-018147e6f498bbc1d
Default VPC No	IPv4 CIDR 10.0.0.0/16	IPv6 pool -	IPv6 CIDR -
Network Address Usage metrics Disabled	Route 53 Resolver DNS Firewall rule groups -	Owner ID 945153889631	

Resource map New

CIDRs

Flow logs

Tags

Resource map Info

10. After selecting create a service role, one has to select the VPC, Subnet and security group options by clicking on them and selecting from the dropdown menu. The image shows the selected ones highlighted in blue.

### Virtual Private Cloud (VPC)

Choose one or more VPCs ▼

VPC\_a3  
vpc-02ebfd3561f5763d6

vpc\_25  
vpc-09dcb33e1bb2cc7ce

### Subnet

Choose one or more subnets ▼

-  
subnet-02cf757853cfa0a95

project-subnet-private1-us-east-2a  
subnet-00f85f08cbc5487e5

### Security group

Choose one or more security groups ▼

ElasticMapReduce-Core  
sg-06bec62fb7aee9c7e

ElasticMapReduce-Primary  
sg-019f76e28b96a4f5d

default  
sg-0fb014c1d0bc86868

11. One can choose an existing instance profile or create an instance profile. If chosen, the option of create an instance profile, the Amazon EMR gives an option of specifying a custom set of



resources for the S3 bucket that will be created later on. It also provides the option of accessing all the buckets or the specific bucket that will be created with read and write access or giving only read access to the bucket

**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.

**S3 bucket access** [Info](#)

☒ Specific S3 buckets or prefixes in your account [Info](#)  
Choose the buckets or prefixes that you want this instance profile to access.

☐ All S3 buckets in this account with read and write access  
Grant the instance profile access to all buckets that have read and write access enabled in your account.

**S3 buckets**  
We've already added the resources that you configured in the **Cluster logs** section. Choose the S3 buckets and bucket prefixes where you store logs and data for your cluster, bootstrap actions, and steps.

S3 URI  
 [View](#) [Browse S3](#) [Add](#)

S3 bucket	Prefix	Permission	
aws-logs-94515388...	elasticmapreduce	Read and write	<a href="#">Edit</a>

Inherited from Cluster logs

**Cluster termination**  
Terminate cluster after idle time  
Idle time: 1 hour

**Cluster logs - optional**

Amazon S3 location  
[s3://aws-logs...](#)

**Identity and Access Management (IAM) roles**

Service role  
New service role

Instance profile  
New instance profile

**Creating cluster**  
Creating instance profile  
0%

[Cancel](#) [Create cluster](#)

12. After completing all the steps, we can click on create cluster, and cluster is created as shown in the image below.

[Amazon EMR](#) > [EMR on EC2: Clusters](#) > My cluster

**My cluster** Updated less than a minute ago [Refresh](#) [Actions](#)

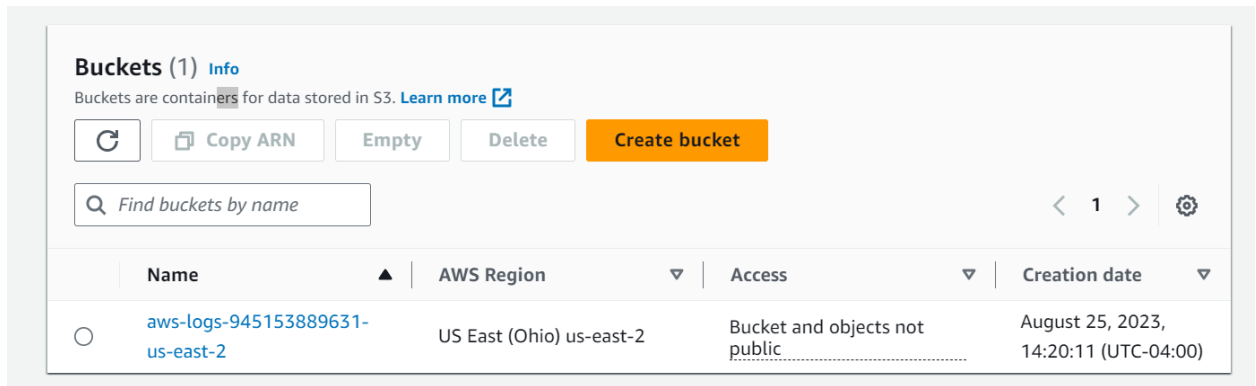
**▼ Summary**

Cluster info	Applications	Cluster management	Status and time
Cluster ID j-OD7OWF7DRIQ3	Amazon EMR version emr-6.12.0	Log destination in Amazon S3 <a href="#">aws-logs-945153889631-us-east-2/elasticmapreduce</a>	Status <span>Starting</span>
Cluster configuration Instance groups	Installed applications Spark 3.4.0, Zeppelin 0.10.1	Primary node public DNS -	Creation time August 25, 2023, 14:20 (UTC-04:00)
Capacity 1 Primary 1 Core 1 Task			Elapsed time -1 seconds

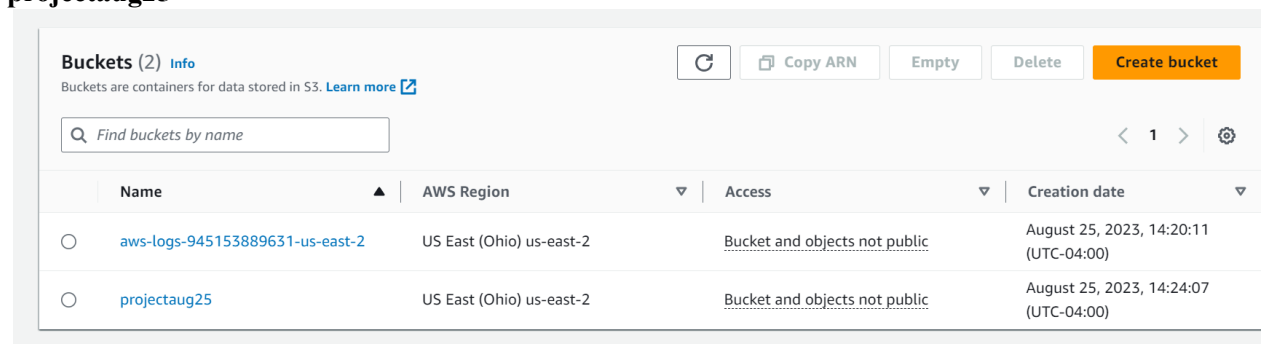
[Properties](#) [Bootstrap actions](#) [Instances \(Hardware\)](#) [Steps](#) [Applications](#) [Configurations](#) [Monitoring](#) [Events](#) [Tags \(1\)](#)

[Operating system](#) [Cluster logs](#) [Cluster termination](#)

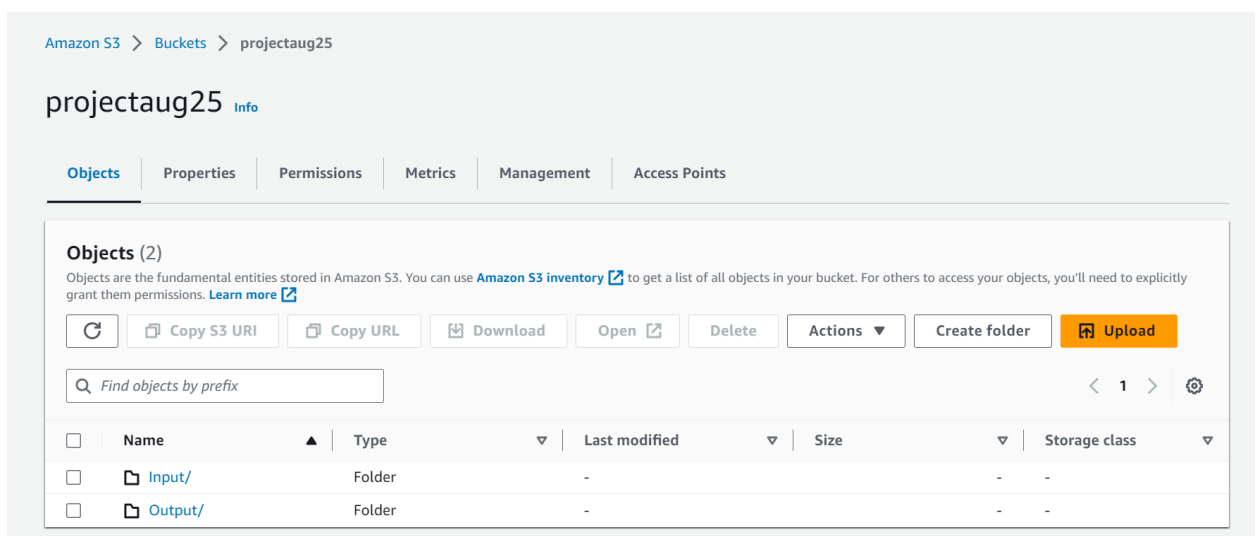
13. Create a bucket now by searching S3 in search bar. It navigates to the page as shown below which shows the existing buckets if present.



14. After clicking create bucket, one can write the name of the bucket and choose the option of giving access of the bucket to the public or not. As shown below, a new bucket is created **projectaug25**



15. After creating the bucket, we have to upload the data files. To upload the files, two folders are created Input and Output folders by clicking on Create folder



16. After clicking on upload button one can Upload the data file (all\_fuels\_data.csv) and python file for spark in Input folder.

Services Search [Alt+S] Global akanakam

**Upload succeeded**  
View details below.

Destination	Succeeded	Failed
s3://projectaug25/Input/	1 file, 2.8 MB (100.00%)	0 files, 0 B (0%)

**Files and folders** Configuration

**Files and folders** (1 Total, 2.8 MB)

Find by name

Name	Folder	Type	Size	Status	Error
all_fuels_data.csv	-	text/csv	2.8 MB	Succeeded	-

17. The spark script written is shown below:

```
from pyspark.sql import SparkSession
from pyspark.ml.feature import VectorAssembler
from pyspark.ml.regression import RandomForestRegressor
from pyspark.ml.evaluation import RegressionEvaluator
from pyspark.sql.functions import col

# Initialize Spark session
spark = SparkSession.builder.appName("FuelPricePrediction").getOrCreate()

# Load data from CSV into a Spark DataFrame
data_path = "s3://projectaug25/Input/all_fuels_data.csv"
data_df = spark.read.csv(data_path, header=True, inferSchema=True)

# Select relevant columns and rename 'close' column to 'label'
selected_data = data_df.select("open", "high", "low", "volume", "close") \
    .withColumnRenamed("close", "label")

# Data Preprocessing and Feature Engineering
assembler = VectorAssembler(inputCols=["open", "high", "low", "volume"], outputCol="features")
assembled_df = assembler.transform(selected_data)

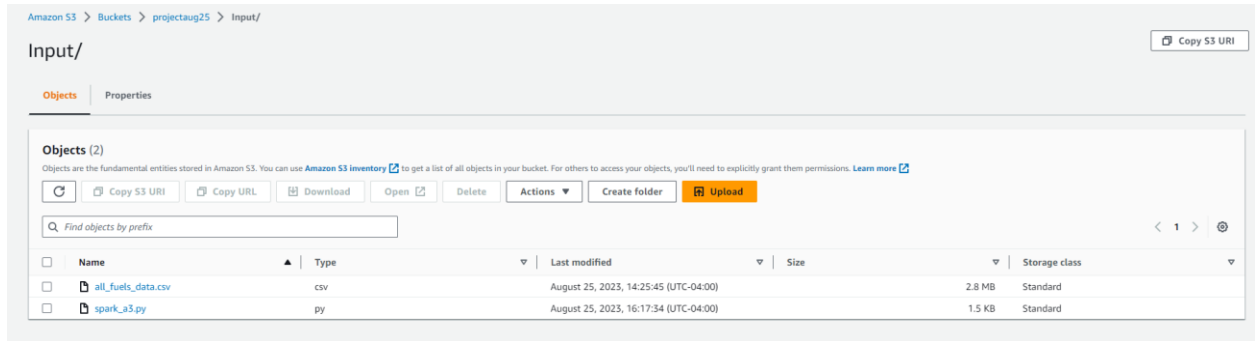
# Split Data into Training and Testing Sets
train_ratio = 0.8
test_ratio = 1.0 - train_ratio
train_data, test_data = assembled_df.randomSplit([train_ratio, test_ratio], seed=12345)

# Build and Train a Machine Learning Model (Random Forest Regressor)
rf_regressor = RandomForestRegressor(featuresCol="features", labelCol="label", numTrees=100)
model = rf_regressor.fit(train_data)

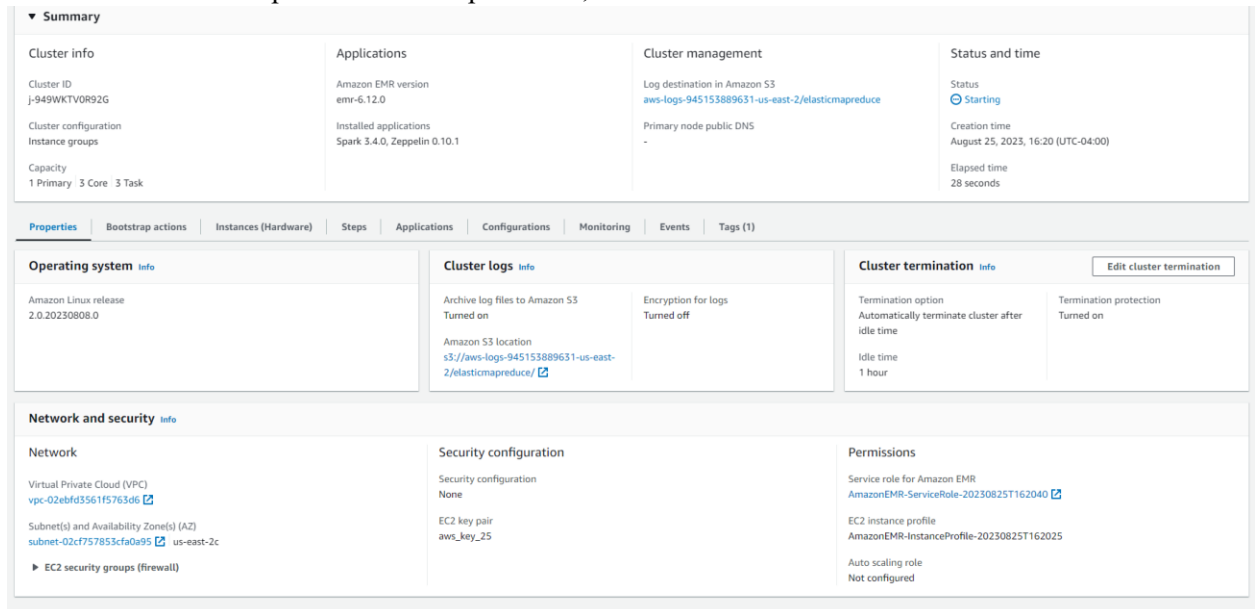
# Make predictions on the test data
predictions = model.transform(test_data)

# Evaluate the model's performance
evaluator = RegressionEvaluator(labelCol="label", predictionCol="prediction", metricName="rmse")
rmse = evaluator.evaluate(predictions)
output_file_path = "s3://projectaug25/Output/output.txt"
with open(output_file_path, "w") as f:
    f.write(f"Root Mean Squared Error (RMSE): {rmse}")
print(f"Root Mean Squared Error (RMSE): {rmse}")
```

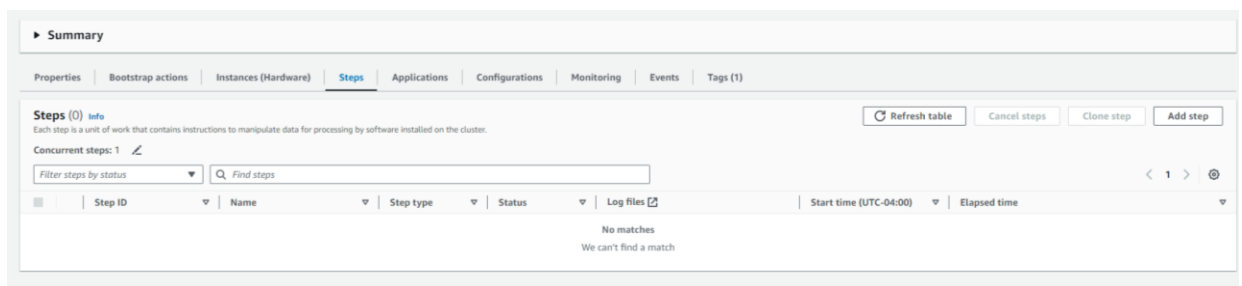
The spark script written is uploaded in the same input folder



18. Once all the files are uploaded in the input folder, Return to Cluster created



19. Go the steps shown in the menu bar



20. Add step, where you can choose the step settings by choosing **Spark application** , and give a name to the step as Spark\_a / SparkAug25 and provide or set the JAR location to spark.py or the python script kept in the bucket.

Amazon EMR > EMR on EC2: Clusters > newAug25 > Add step

## Add step [Info](#)

### Step settings

Type

☒ Custom JAR  
Adds a step that enables you to write a custom script to process your data using the Java programming language.

☐ Streaming program  
Adds a step that uses standard input to run mapper/reducer scripts and send results to standard output.

☐ Spark application  
Adds a step that submits work to the Spark framework on the cluster.

☐ Shell script  
Troubleshoot your cluster.

Name

JAR location

The JAR location may be a path into S3 or a fully qualified java class in the classpath.

Arguments - *optional* [Info](#)

These are passed to the main function in the JAR. If the JAR does not specify a main class in its manifest file, you can specify another class name as the first argument.

Choose Amazon S3 location

S3 buckets > projectaug25 > Input

Objects (1/2)

Find objects

Key

all\_fuel\_data.csv

spark\_a3.py

Cancel Choose

21. The step starts the process of running the script. Once completed its shown in the status as shown in the picture.

Summary

Properties | Bootstrap actions | Instances (Hardware) | **Steps** | Applications | Configurations | Monitoring | Events | Tags (1)

Steps (1) [Info](#)


Each step is a unit of work that contains instructions to manipulate data for processing by software installed on the cluster.

Concurrent steps: 1

Filter steps by status Find steps

	Step ID	Name	Step type	Status	Log files	Start time (UTC-04:00)	Elapsed time
<input type="checkbox"/>	s-01442203D4XJE0DJX...	Spark_a	Spark submit	Completed	controller: syslog stderr: stdout	August 25, 2023 at 17:47	1 minute, 4 seconds

22. The details regarding the spark session can be seen in the history of the server.



3.3.1-amzn-0

History Server

Event log directory: s3a://prod-us-east-1-appinfo-arcj-1DGMS6DTUR34M/sparklogs

Last updated: 2023-08-25 18:01:17

Client local time zone: America/New\_York

Search:

Version	App ID	App Name	Started	Completed	Duration	Spark User	Last Updated	Event Log
3.4.0-amzn-0	application_169299999202_0001	FuelPricePrediction	2023-08-25 17:48:11	2023-08-25 17:48:37	26 s	hadoop	2023-08-25 17:48:58	<a href="#">Download</a>

Showing 1 to 1 of 1 entries

[Show incomplete applications](#)

23. One can click on Application Id and look into the completed jobs and tasks that are completed for running the spark application

Spark Jobs <sup>(?)</sup>

User: hadoop

Total Uptime: 26 s

Scheduling Mode: FIFO

Completed Jobs: 11

Event Timeline

Completed Jobs (11)

Page: 1

1 Pages. Jump to 1. Show 100 items in a page. Go

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
10	treeAggregate at Statistics.scala:58 treeAggregate at Statistics.scala:58	2023/08/25 21:48:36	0.7 s	1/1	1/1
9	collectAsMap at RandomForest.scala:663 collectAsMap at RandomForest.scala:663	2023/08/25 21:48:35	0.9 s	2/2	2/2
8	collectAsMap at RandomForest.scala:663 collectAsMap at RandomForest.scala:663	2023/08/25 21:48:34	0.6 s	2/2	2/2
7	collectAsMap at RandomForest.scala:663 collectAsMap at RandomForest.scala:663	2023/08/25 21:48:34	0.5 s	2/2	2/2
6	collectAsMap at RandomForest.scala:663 collectAsMap at RandomForest.scala:663	2023/08/25 21:48:33	0.6 s	2/2	2/2
5	collectAsMap at RandomForest.scala:663 collectAsMap at RandomForest.scala:663	2023/08/25 21:48:32	1 s	2/2	2/2
4	collectAsMap at RandomForest.scala:1054 collectAsMap at RandomForest.scala:1054	2023/08/25 21:48:31	1 s	2/2	2/2
3	aggregate at DecisionTreeMetadata.scala:125 aggregate at DecisionTreeMetadata.scala:125	2023/08/25 21:48:30	0.8 s	1/1	1/1
2	take at DecisionTreeMetadata.scala:119 take at DecisionTreeMetadata.scala:119	2023/08/25 21:48:28	2 s	1/1	1/1
1	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2023/08/25 21:48:25	2 s	1/1	1/1
0	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2023/08/25 21:48:21	4 s	1/1	1/1

Page: 1

1 Pages. Jump to 1. Show 100 items in a page. Go

24. The output can be shown in the output folder or the additional log files

Amazon S3

>

Buckets

>

projectaug25

>

Output/

Output/

Objects

Properties

Objects (1)

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 yo

↻

Copy S3 URI

Copy URL

Download

Open


Delete

Actions ▼

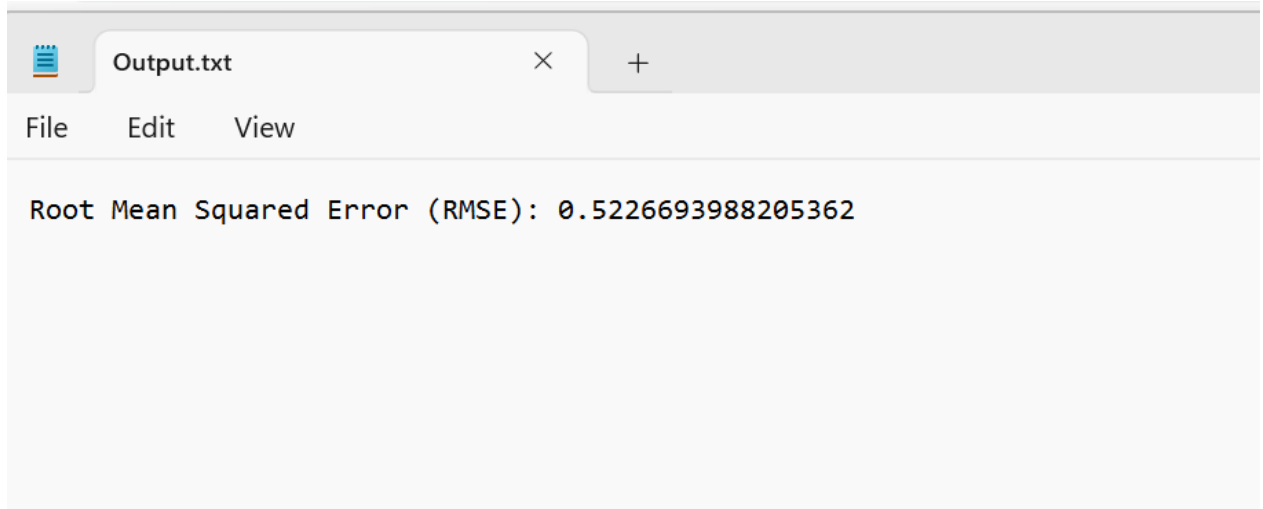
Create

Find objects by prefix

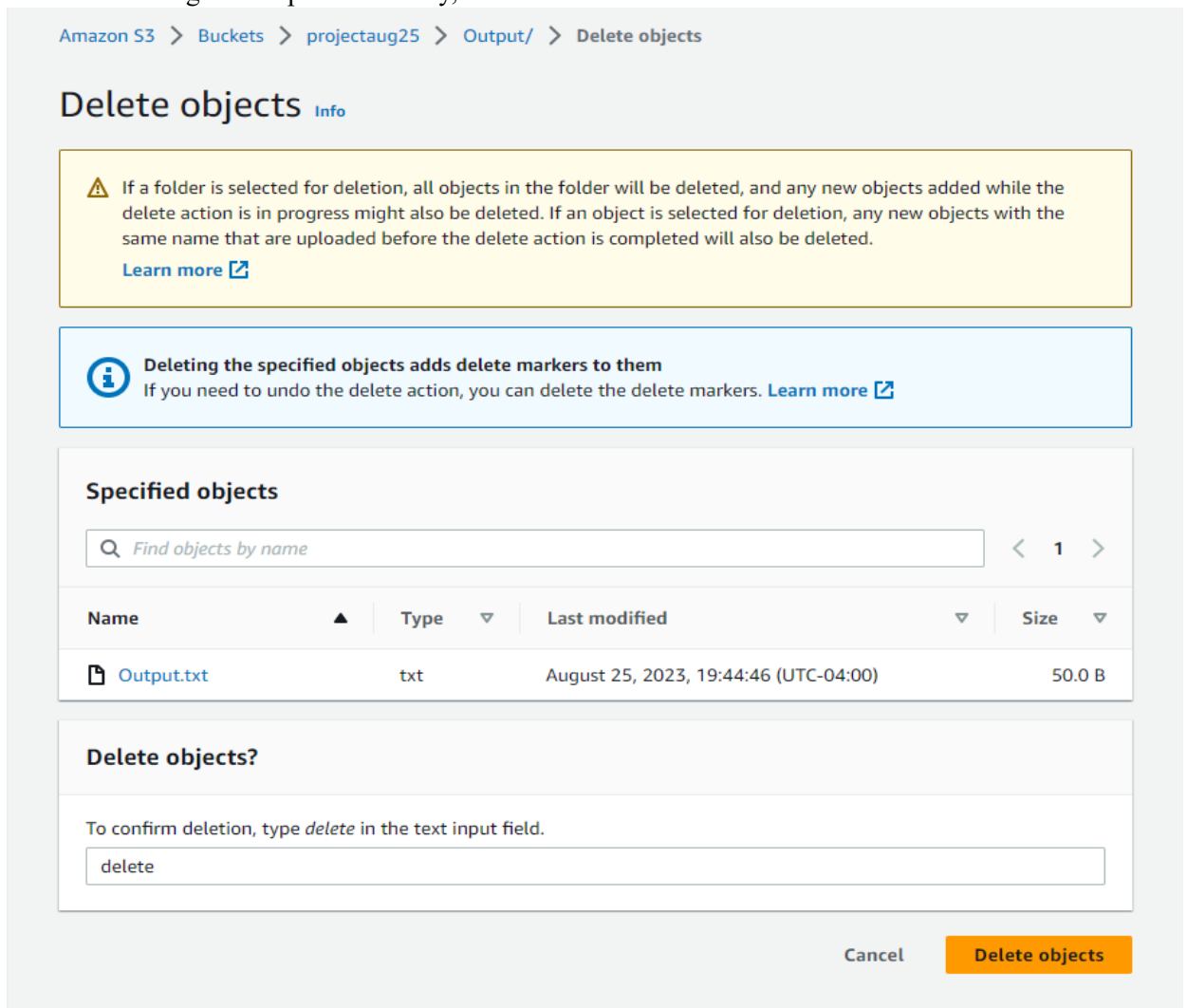
Show versions

<input type="checkbox"/>	Name	Type	Last modified
<input type="checkbox"/>	 Output.txt	txt	August 25, 2023, 1

25. The output of the machine learning script is shown below




26. After the running the script successfully, delete the files in the bucket and delete the bucket.



Amazon S3 > Buckets > projectaug25 > Delete bucket

## Delete bucket [Info](#)



- Deleting a bucket cannot be undone.
- Bucket names are unique. If you delete a bucket, another AWS user can use the name.
- If this bucket is used with a Multi-Region Access Point in an external account, initiate failover before deleting the bucket.
- If this bucket is used with an access point in an external account, the requests made through those access points will fail after you delete this bucket.


[Learn more](#)

### Delete bucket "projectaug25"?

To confirm deletion, enter the name of the bucket in the text input field.

[Cancel](#)
[Delete bucket](#)

27. Terminate the cluster also to avoid getting charged for using the idle amazon web service

Updated less than a minute ago 

**Actions** ▲

- Clone cluster
- View command for cloning cluster
- Terminate cluster

#### Applications


Amazon EMR version  
emr-6.12.0

Installed applications  
Spark 3.4.0, Zeppelin 0.10.1


#### Cluster management

Log destination in Amazon S3  
Logging not configured

Persistent application UIs  
[Spark History Server](#)  
[YARN timeline server](#)

Primary node public DNS  
 ec2-3-131-169-76.us-east-2.compute.amazonaws.com  
[Connect to the Primary Node using SSH](#)



#### Status and time



Status  
 **Waiting**






Creation time  
August 25, 2023, 16:49 (UTC-04:00)

Elapsed time  
3 hours, 6 minutes

**Clusters (5)** [Info](#)

Filter clusters by status    View details [Terminate](#) [Clone](#) [Create cluster](#)

Filter clusters by creation date-time  < 1 > 

	Cluster Id	Cluster name	Status	Creation time (UTC-04:00)	Elapsed time	Normalized instance hours
<input type="checkbox"/>	<a href="#">j-2PUJGW2J4M5F8</a>	newCluster25	 <b>Terminating</b> User request	August 25, 2023, 16:49	3 hours, 7 minutes	224
<input type="checkbox"/>	<a href="#">j-949WKTVO92G</a>	newAug25	 <b>Terminated</b> User request	August 25, 2023, 16:20	28 minutes, 24 seconds	56
<input type="checkbox"/>	<a href="#">j-QD7QWF7DRIQ3</a>	My cluster	 <b>Terminated</b> Auto-terminate	August 25, 2023, 14:20	1 hour, 10 minutes	48
<input type="checkbox"/>	<a href="#">j-36SIGXMIUPLKKK</a>	My cluster	 <b>Terminated</b> User request	August 22, 2023, 18:25	34 minutes, 53 seconds	56
<input type="checkbox"/>	<a href="#">j-KOIEY6SKLMA</a>	My cluster	 <b>Terminated</b> User request	August 09, 2023, 23:46	1 hour, 36 minutes	112

28. As shown the cluster is terminated



Clusters (5) 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	Normalized instance hours
<input type="checkbox"/>	<a href="#">j-2PUDGW2J4M5F8</a>	newCluster25	Terminated User request	August 25, 2023, 16:49	3 hours, 8 minutes	224
<input type="checkbox"/>	<a href="#">j-949WKTVO92G</a>	newAug25	Terminated User request	August 25, 2023, 16:20	28 minutes, 24 seconds	56
<input type="checkbox"/>	<a href="#">j-0D7QWF7DRIO3</a>	My cluster	Terminated Auto-terminate	August 25, 2023, 14:20	1 hour, 10 minutes	48
<input type="checkbox"/>	<a href="#">j-36SIGXMUPLKKK</a>	My cluster	Terminated User request	August 22, 2023, 18:25	34 minutes, 53 seconds	56
<input type="checkbox"/>	<a href="#">j-KOIEY6SKLMA</a>	My cluster	Terminated User request	August 09, 2023, 23:46	1 hour, 36 minutes	112

## CORE COMPONENTS OF THE PROJECT

1. As mentioned the data needs to have more than 1000 rows, the data we got from Kaggle contains more than 27000 rows. The csv file is chosen as data lake as it presents structured data and easy to handle.
2. The data lake is uploaded or connected to AWS distributed cloud services. The choice of aws is primarily because of
  - The familiarity with the interface during the coursework and assignments given.
  - AWS has greater resources, infrastructure and superior, scalable services than Azure.
  - The pricing model of AWS is on hourly basis.
  - It has rich and user friendly interface.
3. The objective was to predict the closing prices of fuel data by utilizing daily attributes such as fuel rates, consumption, and the high and low costs on the following day. This involved training the data and subsequently assessing the predictive capability. The evaluation encompassed computing the root mean square error or accuracy of the chosen model. The models under consideration included the random forest regressor as well as other machine learning approaches like linear regression. At the moment, the focus is on showcasing a single model, with the resultant output being stored within a text file housed in the S3 bucket.

All the step-by-step procedures have been mentioned in the file with pictures.

Thankyou.