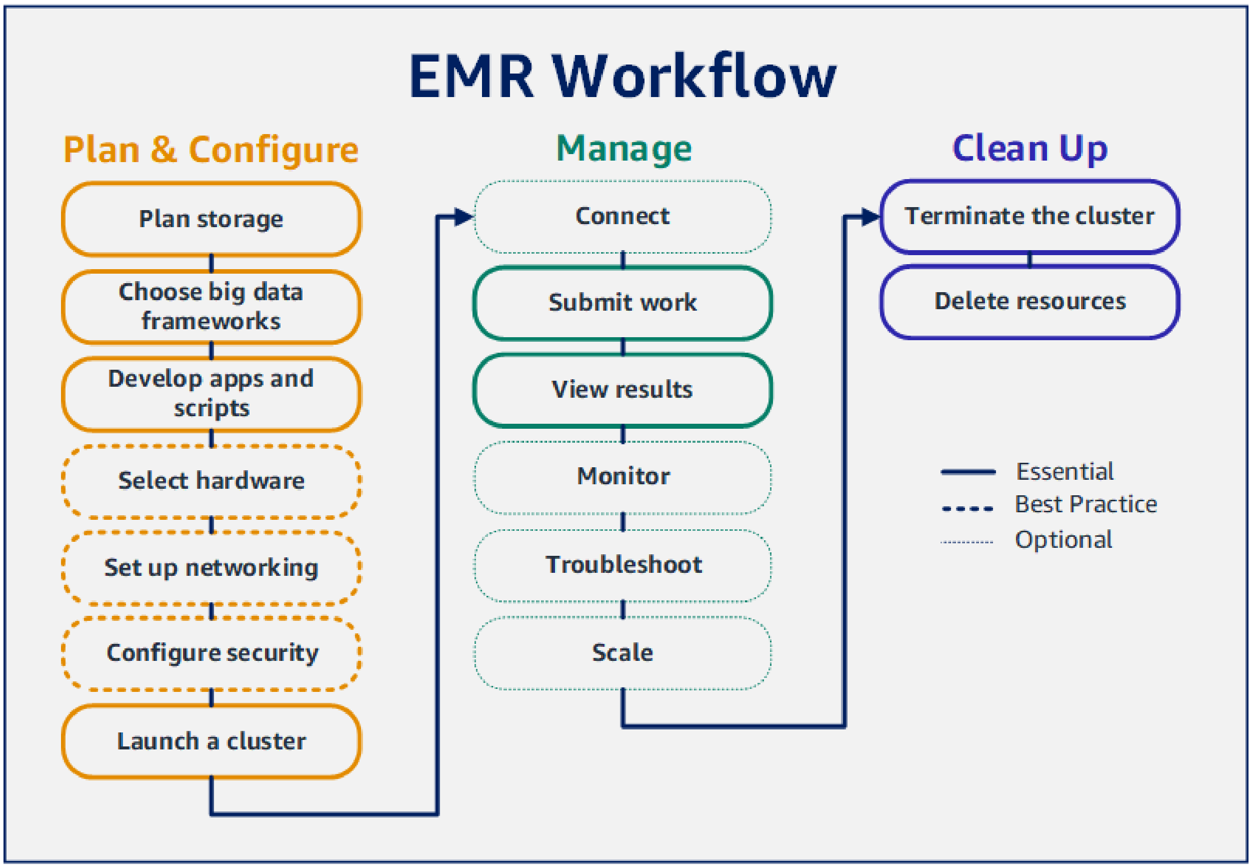
Experiment with advanced data analytics techniques using cloud-based big data services (e.g., AWS EMR, Azure HDInsight, Google Dataproc).

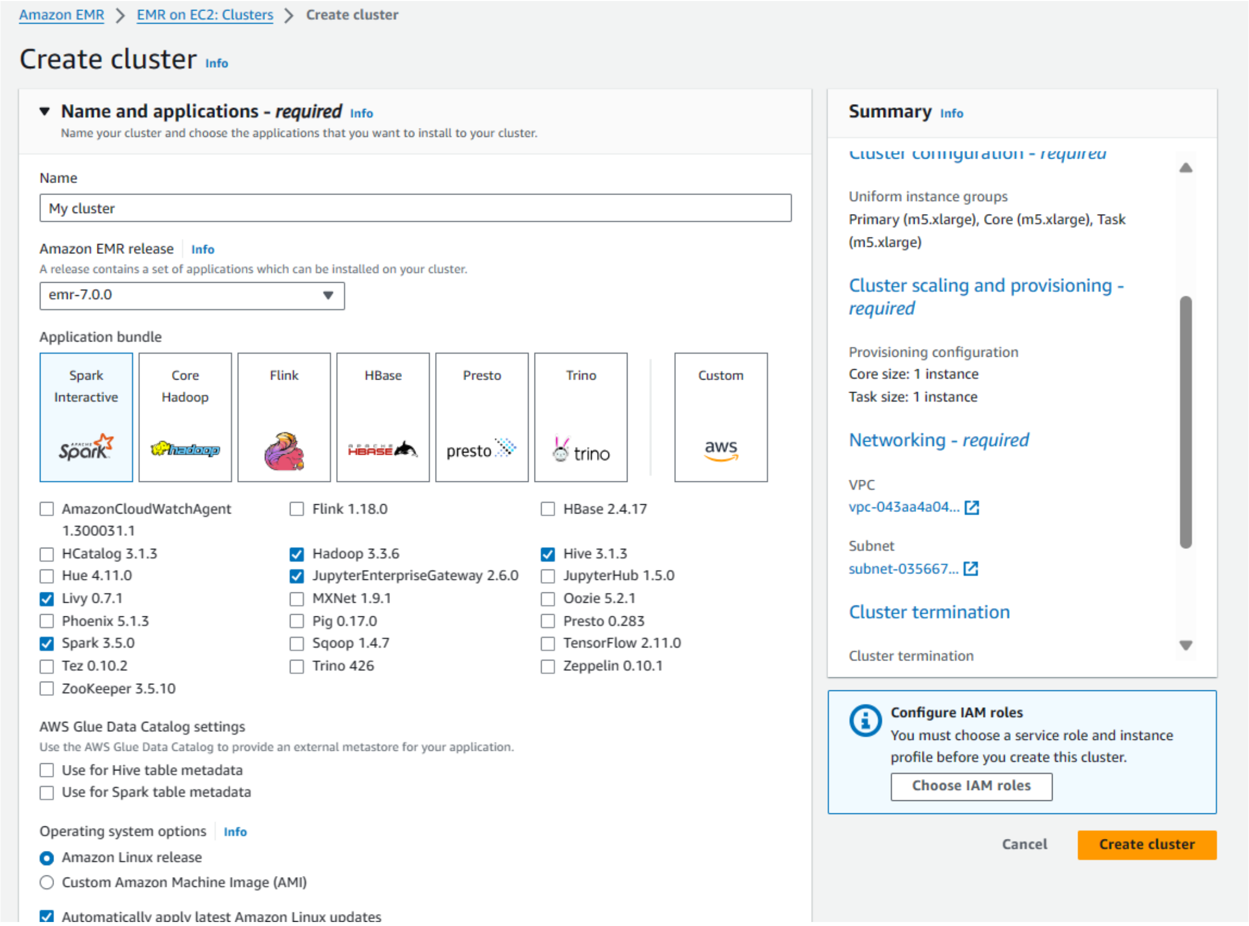
Ebuka Obiakor – 16th March 2024

What is Amazon EMR?

Amazon EMR (previously called Amazon Elastic MapReduce) is a managed cluster platform that simplifies running big data frameworks, such as Apache Hadoop and Apache Spark, on AWS to process and analyze vast amounts of data.

Tutorial followed: Tutorial: Getting started with Amazon EMR - Amazon EMR





Create cluster

▼ Name and applications - *required*

Name your cluster and choose the applications that you want to install to your cluster.

Name

My cluster

Amazon EMR release

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

emr-7.0.0

Application bundle

Spark Interactive | Core Hadoop | Flink | HBase | Presto | Trino | Custom

- ☐ AmazonCloudWatchAgent 1.300031.1
- ☐ HCatalog 3.1.3
- ☐ Hue 4.11.0
- ☑ Livy 0.7.1
- ☐ Phoenix 5.1.3
- ☑ Spark 3.5.0
- ☐ Tez 0.10.2
- ☐ ZooKeeper 3.5.10
- ☐ Flink 1.18.0
- ☑ Hadoop 3.3.6
- ☑ JupyterEnterpriseGateway 2.6.0
- ☐ MXNet 1.9.1
- ☐ Pig 0.17.0
- ☐ Sqoop 1.4.7
- ☐ Trino 426
- ☐ HBase 2.4.17
- ☑ Hive 3.1.3
- ☐ JupyterHub 1.5.0
- ☐ Oozie 5.2.1
- ☐ Presto 0.283
- ☐ TensorFlow 2.11.0
- ☐ 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 Hive table metadata
- ☐ Use for Spark table metadata

Operating system options

- ● Amazon Linux release
- ○ Custom Amazon Machine Image (AMI)

☑ Automatically apply latest Amazon Linux updates

Summary

Cluster configuration - *required*

Uniform instance groups

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

Cluster scaling and provisioning - *required*

Provisioning configuration

Core size: 1 instance

Task size: 1 instance

Networking - *required*

VPC

vpc-043aa4a04...

Subnet

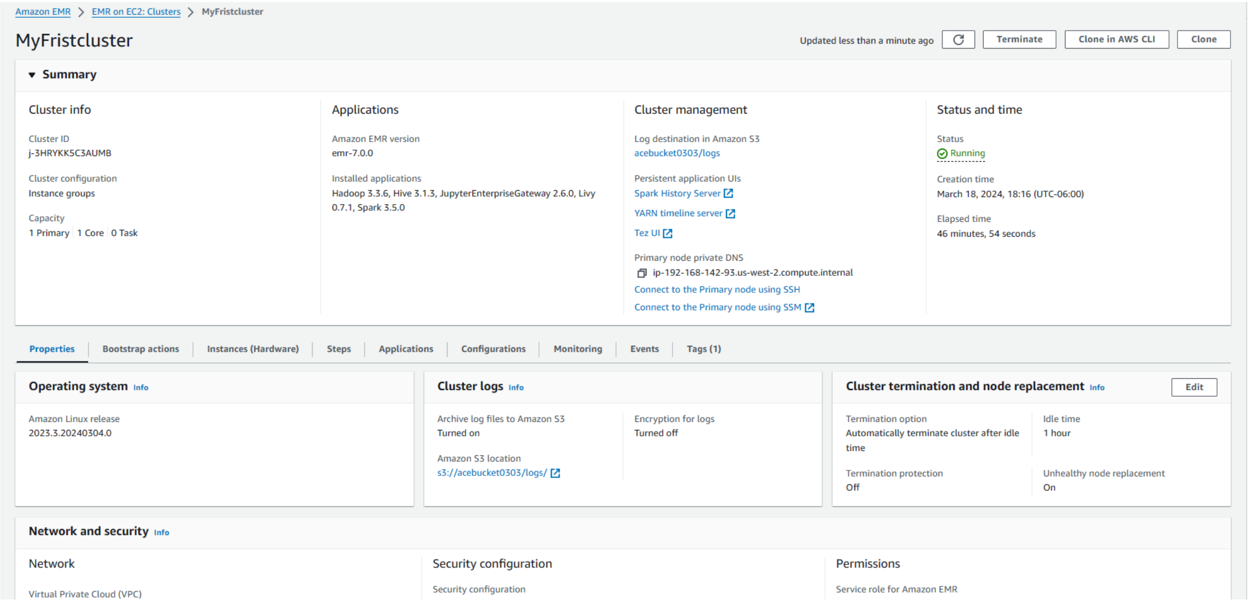
subnet-035667...

Cluster termination

Cluster termination

Configure IAM roles

You must choose a service role and instance profile before you create this cluster.



MyFristcluster

Updated less than a minute ago

▼ Summary

Cluster info

Cluster ID

j-3HRYKK5C3AUMB

Cluster configuration

Instance groups

Capacity

1 Primary | 1 Core | 0 Task

Applications

Amazon EMR version

emr-7.0.0

Installed applications

Hadoop 3.3.6, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.7.1, Spark 3.5.0

Cluster management

Log destination in Amazon S3

acebucket0303/logs

Persistent application UIs

Spark History Server

YARN timeline server

Tez UI

Primary node private DNS

ip-192-168-142-93.us-west-2.compute.internal

Connect to the Primary node using SSH

Connect to the Primary node using SSM

Status and time

Status

Running

Creation time

March 18, 2024, 18:16 (UTC-06:00)

Elapsed time

46 minutes, 54 seconds

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

Operating system

Amazon Linux release

2023.3.20240304.0

Cluster logs

Archive log files to Amazon S3

Turned on

Amazon S3 location

s3://acebucket0303/logs/

Encryption for logs

Turned off

Cluster termination and node replacement

Termination option

Automatically terminate cluster after idle time

Idle time

1 hour

Termination protection

Off

Unhealthy node replacement

On

Network and security

Network

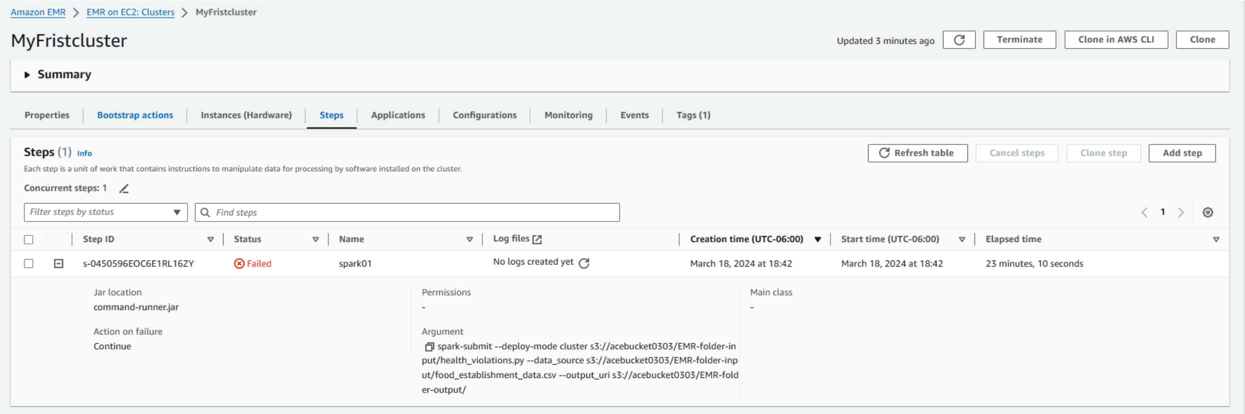
Virtual Private Cloud (VPC)

Security configuration

Security configuration

Permissions

Service role for Amazon EMR



MyFristcluster

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

	Step ID	Status	Name	Log files	Creation time (UTC-06:00)	Start time (UTC-06:00)	Elapsed time
□	s-0450596EOC6E1RL16ZY	Failed	spark01	No logs created yet	March 18, 2024 at 18:42	March 18, 2024 at 18:42	23 minutes, 10 seconds

Jar location
command-runner.jar

Action on failure
Continue

Permissions
-

Argument

```
spark-submit --deploy-mode cluster s3://acebucket0303/EMR-folder-input/health_violations.py --data_source s3://acebucket0303/EMR-folder-input/food_establishment_data.csv --output_uri s3://acebucket0303/EMR-folder-output/
```

Main class
-

Step failed to run, will debug later ...lol ;)

References

https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-gs.html