Launch an EMR cluster- Steps 1-4

Create service linked role for Spot using Cloud9

aws iam create-service-linked-role --aws-service-name spot.amazonaws.com

Upload spark script and dataset to S3 bucket in AWS Console

- Upload <u>Spark Script</u> to S3 application bucket: spark-app-<account-id>
- Create a new S3 folder: spark-app-<account-id>/input
- Upload <u>Dataset</u> to Input folder

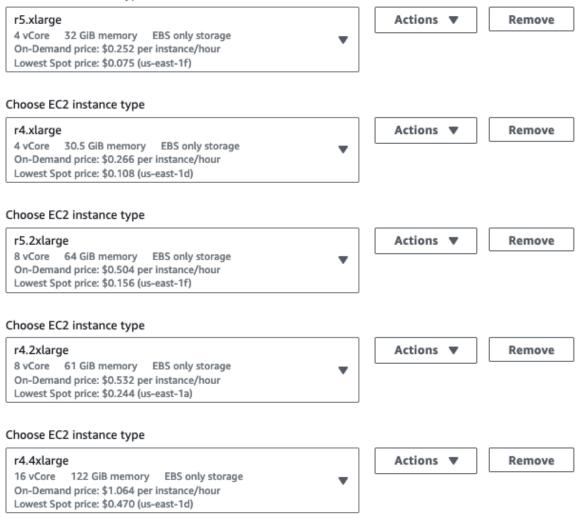
Cluster Configuration

- Select Instance Fleets
- For **Primary**, click on **Add instance type** and add additional instances: m4.xlarge, m5a.xlarge, m5d.xlarge
- For **Core**, click on **Add instance type** and add additional instances recommended in previous section using ec2-instance-select tool. **Note**: change the default instance type already populated (m5.xlarge, which only has 4vCPU and 16GB memory)

Core

Choose one or more EC2 instance type

Choose EC2 instance type

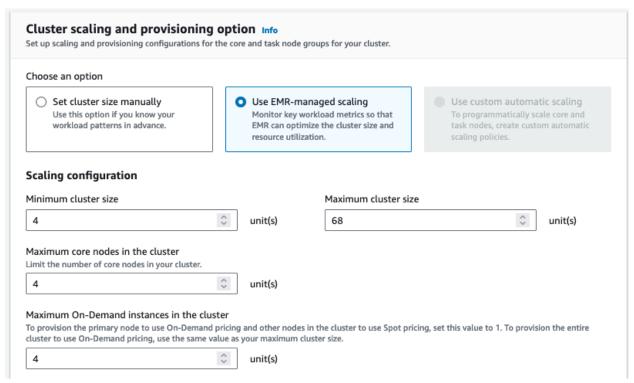


• For **Task**, click on **Add Instance Type**. Add up to 15 instance types based on instances recommended in previous section. **Note**: change the default instance type already populated (m5.xlarge, which only has 4vCPU and 16GB memory)

Cluster scaling and provisioning option as shown in Figure below.

Scaling configuration

- Select Use EMR-managed scaling
- Set minimum cluster size to 4 units
- Set maximum cluster size to 68 units
- Set maximum core nodes to 4 units
- Set maximum On-demand instances in the cluster to 4 units
- Click on the checkbox Apply allocation strategy



Provisioning configuration

- Set Core On-Demand size to 4 units. Leave Core spot size at 0 units
- Set Task Spot size to 32 units.
- Leave timeout configuration settings default.

Provisioning configuration

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



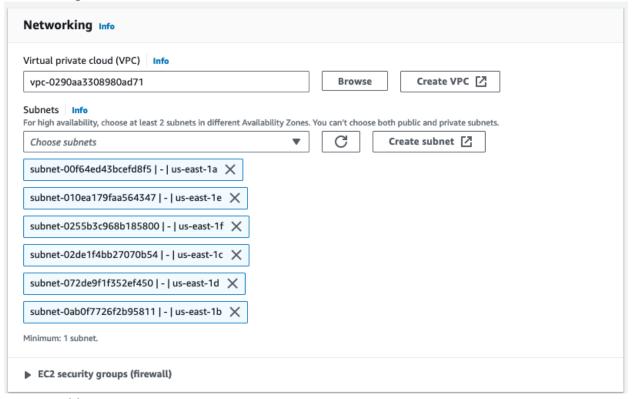
Allocation Strategy

Spot Strategy: Choose Capacity optimized to minimize risk of interruption

Apply allocation strategy (recommended) The allocation strategy determines which of your available pools to request Spot Instances from. Amazon EMR always provisions On-Demand capacity with the lowest-price strategy. On-Demand strategy Lowest price Spot strategy Capacity optimized Price-capacity optimized (recommended) Request the lowest priced Spot Instances from your most Request Spot Instances from your most available pools. This available pools. This is the best strategy to balance instance price strategy has the lowest risk of interruption. and the risk of interruption. Diversified across all pools Lowest price Request Spot instances from the lowest priced pools based on Request Spot Instances evenly across all your available pools. your instance type requirements. This strategy has the highest risk of interruption.

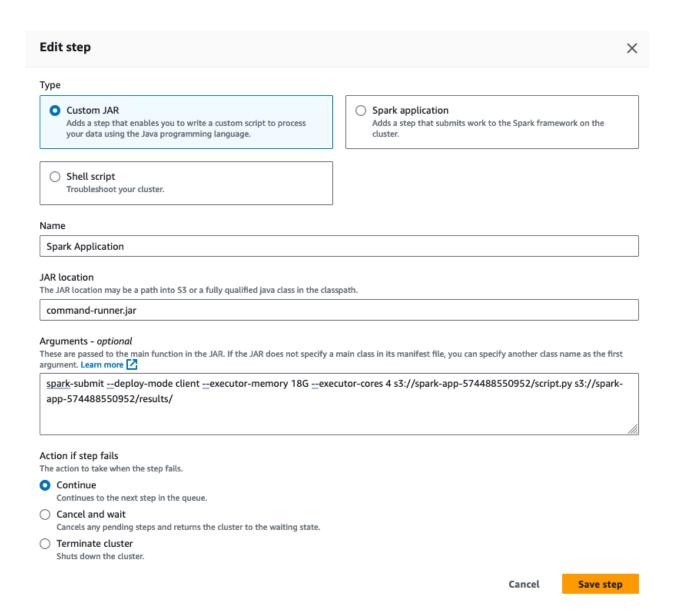
Networking

Allocation strategy Info



Steps - Add

- Configure Steps as show in figure below
- Arguments: spark-submit --executor-memory 18G --executor-cores 4 s3://spark-app-574488550952/script.py s3://spark-app-574488550952/input s3://spark-app-574488550952/results



Skip the following sections: Cluster Termination, Bootstrap actions, Cluster Logs

For Tags - optional, click on **Add new tag**. For **Key**, enter Name. For **Value**, enter EMRTransientCluster1

Leave Software settings default.

Security configuration and EC2 key pair

On the EC2 key pair drop-down, select emr-workshop-key-pair

Identity and Access Management (IAM) roles

- For Amazon EMR service role, select Create a service role
- For Security group, select default

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.

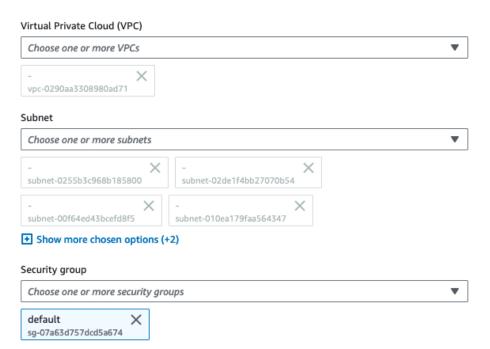
Choose an existing service role
Select a default service role or a custom role with IAM policies attached so that your cluster can interact with other AWS services.

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.

Networking resources

We've already added the resources that you configured in the **Networking** section. Choose the VPC, subnet, and security groups that the service role can access.



EC2 instance profile for Amazon EMR

- Select Create an instance profile
- For S3 bucket access, select All S3 buckets in this account with read and write access