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

Cluster Configuration

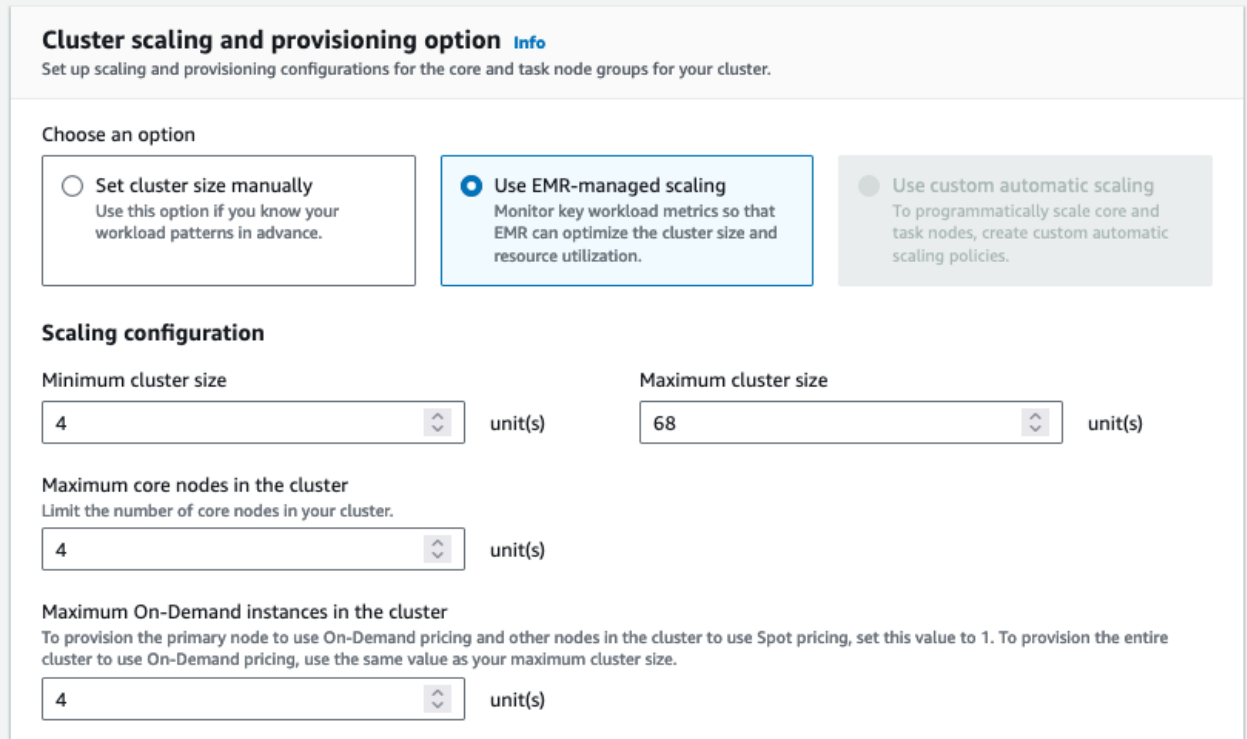
- Select **Instance Fleets**
- For **Primary,** click on **Add instance type** and add additional instances: m4.xlarge, m5a.xlarge, m5d.xlarge
- For **Core**, click on **Add instance type** and add additional instances recommended in previous section using ec2-instance-select tool. **Note**: change the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

- 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.

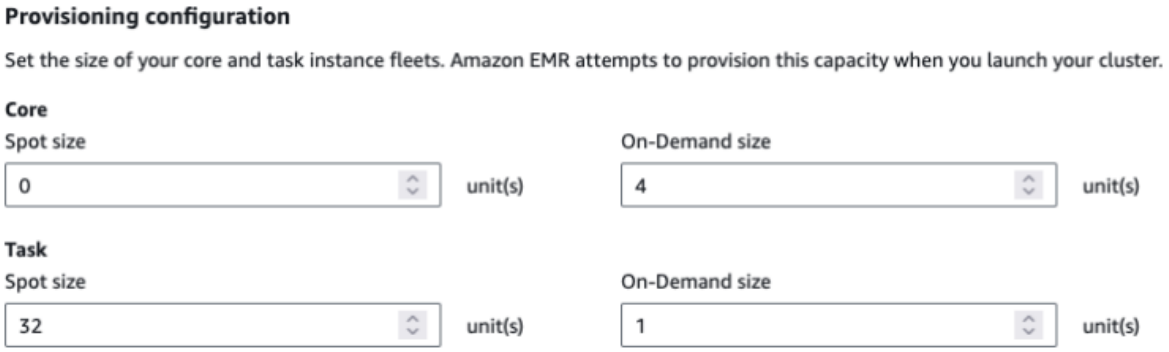
Scaling configuration

- 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**



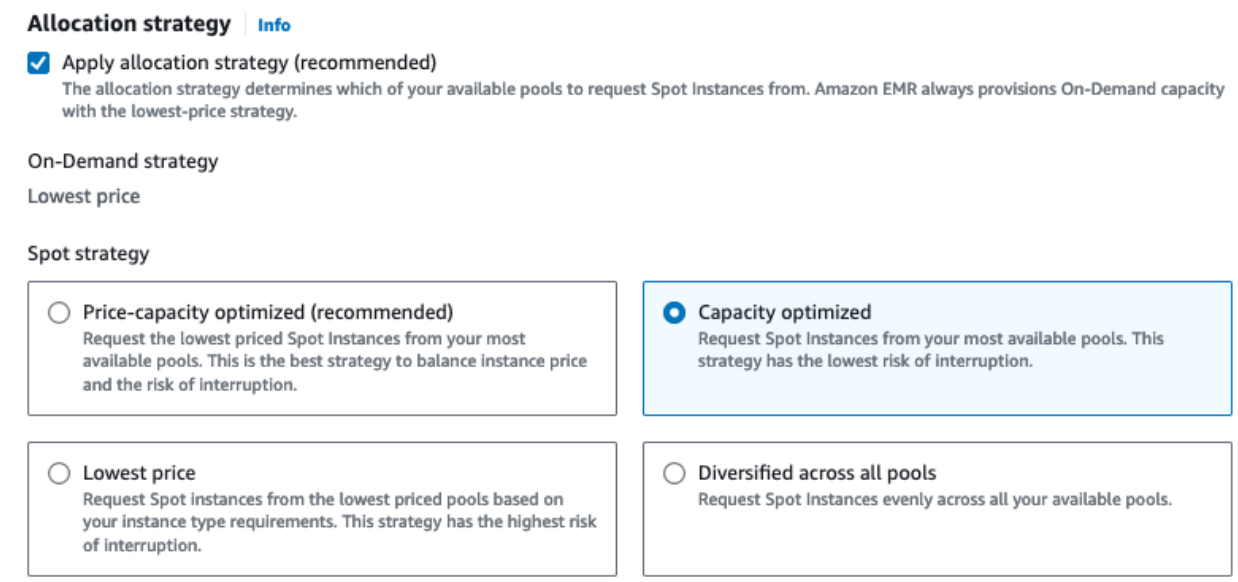
Provisioning configuration

- 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.

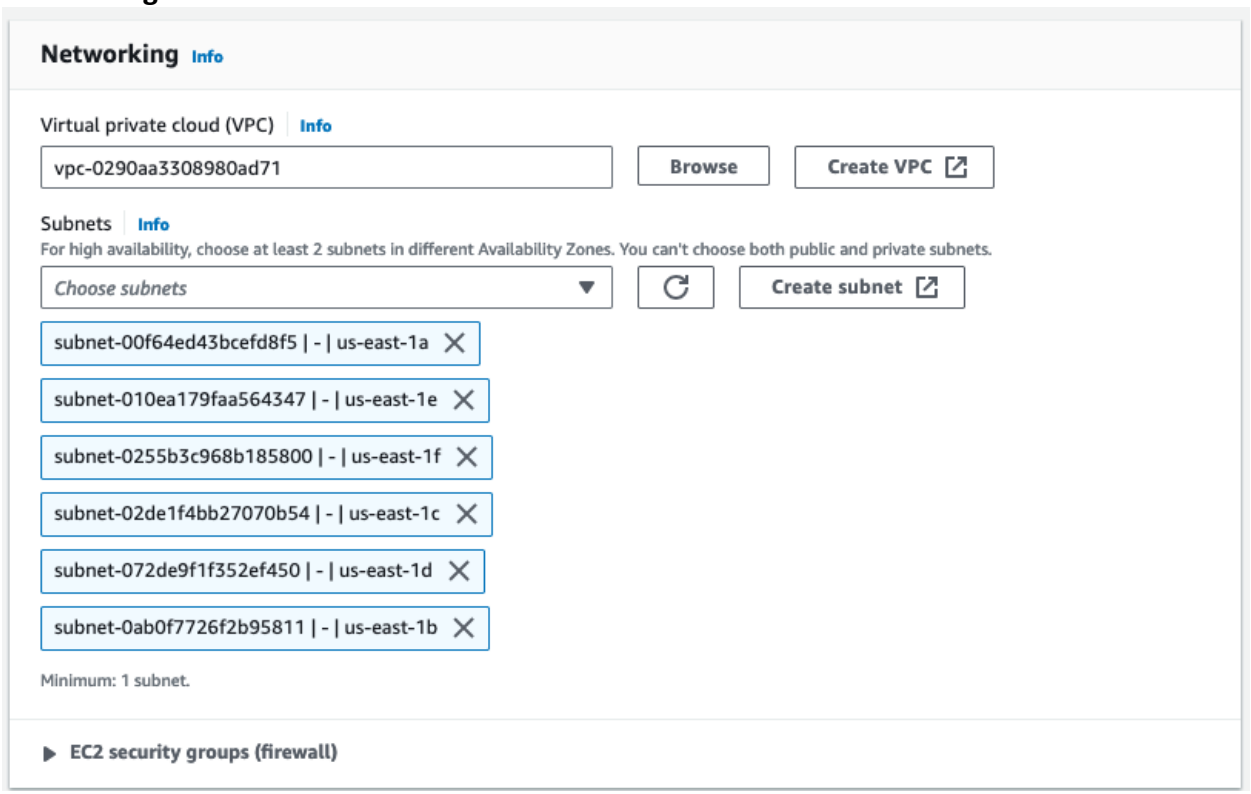


Allocation Strategy

- Spot Strategy: Choose Capacity optimized to minimize risk of interruption

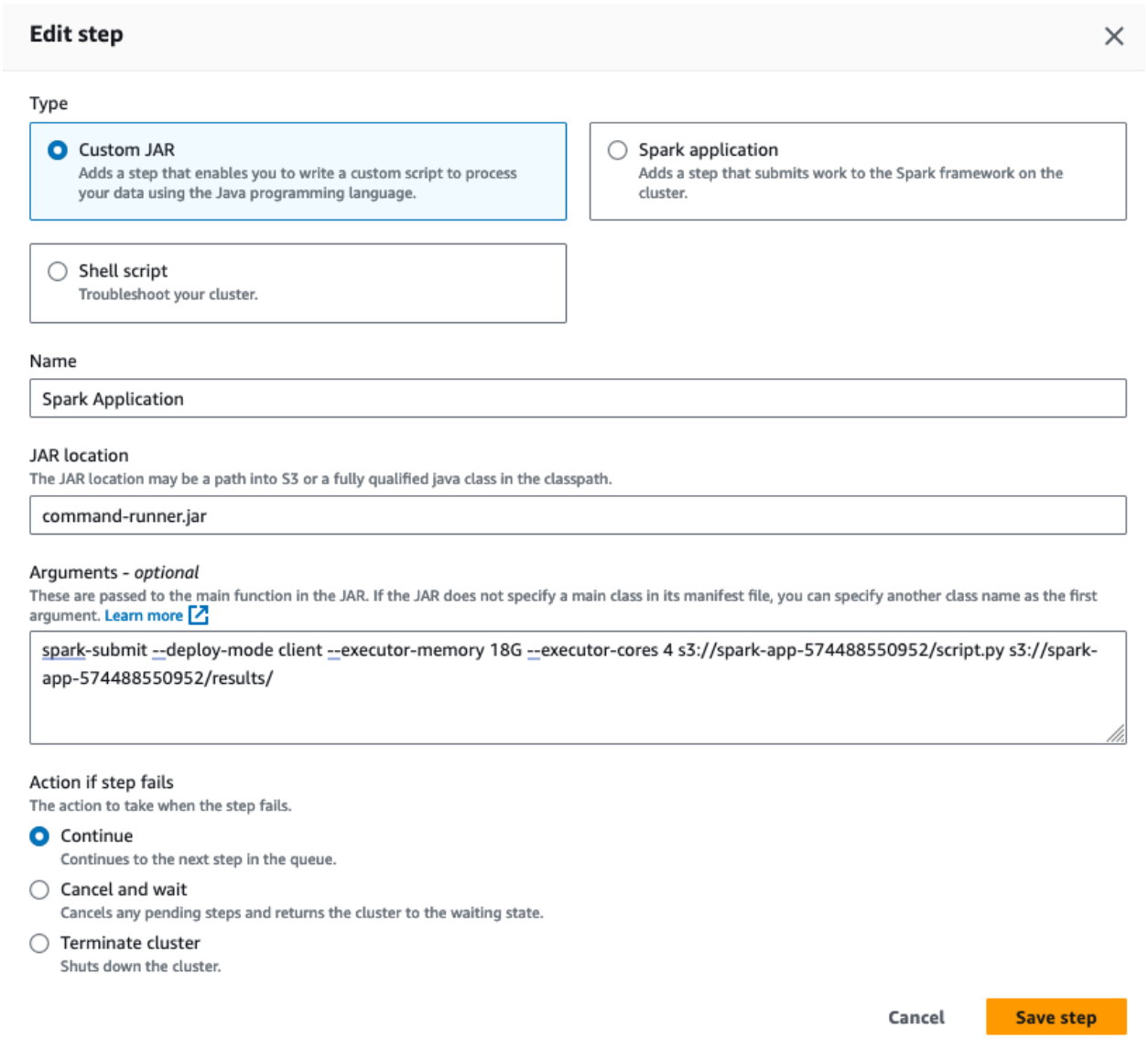


Networking



Steps - *Add*

- Configure Steps as show in figure below
- **Arguments:** spark-submit —deploy-mode client —executor-memory 18G —executor-cores 4 s3://spark-app-574488550952/script.py 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

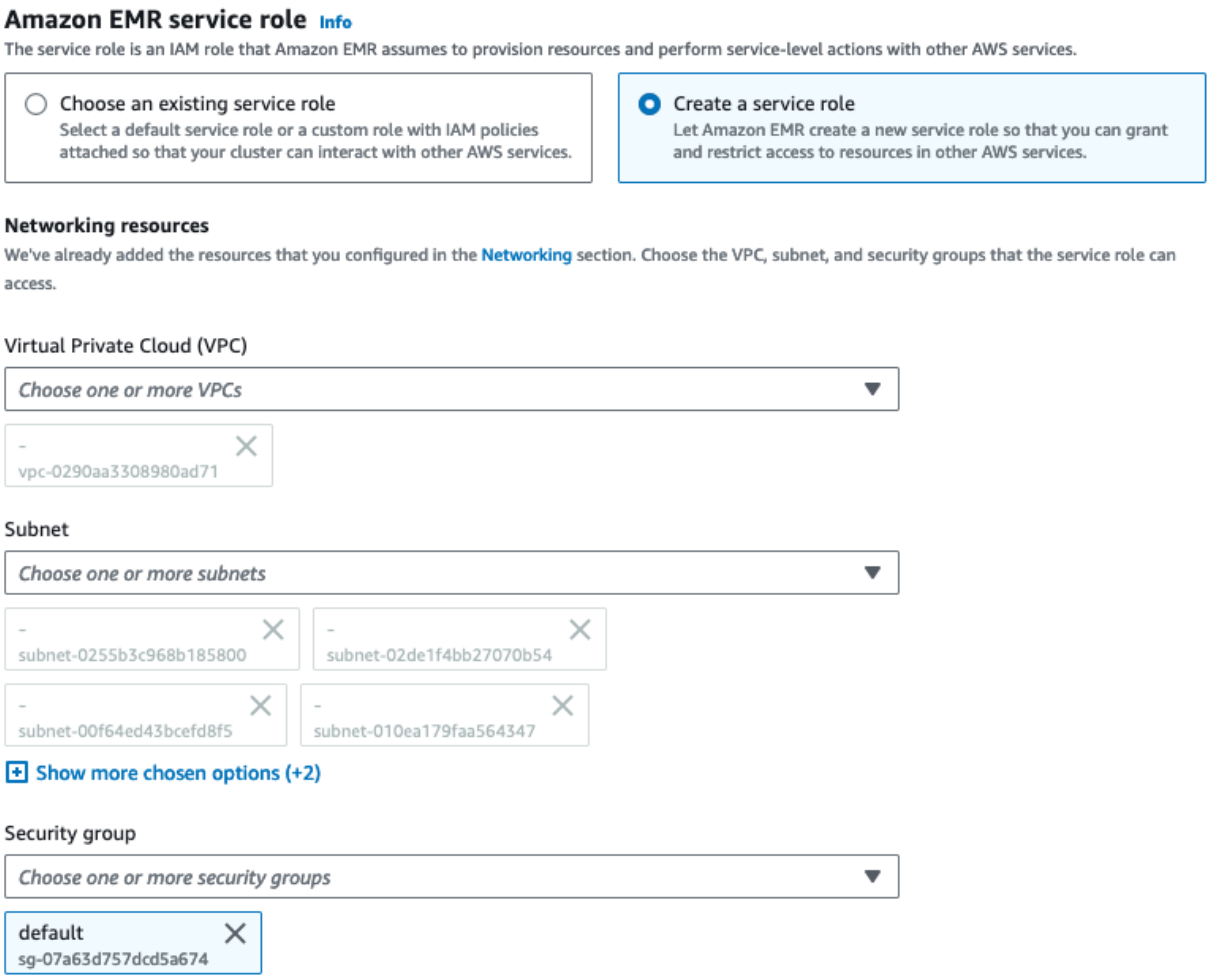
Leave **Software settings** default.

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



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**