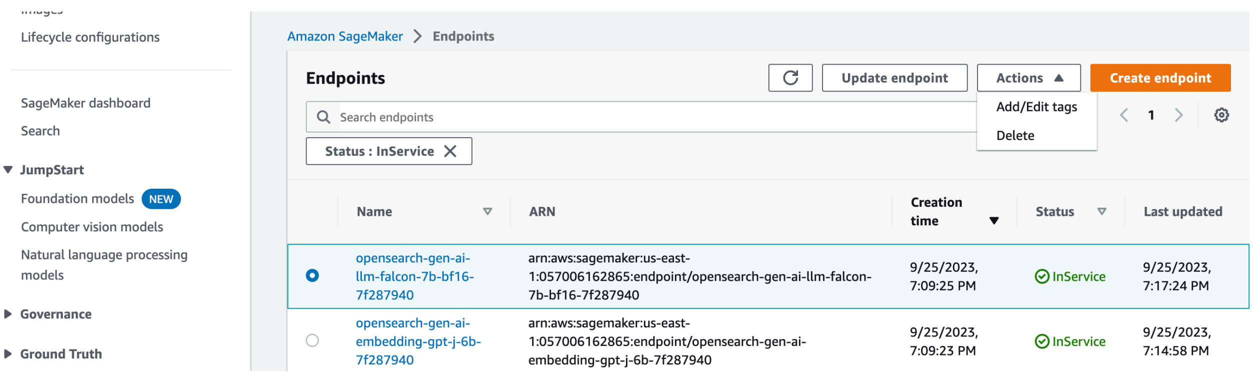
Lab 3: RAG application using Amazon OpenSearch

1. Complete the pre-requisites for the lab
    a. Delete both the endpoints created in the previous labs

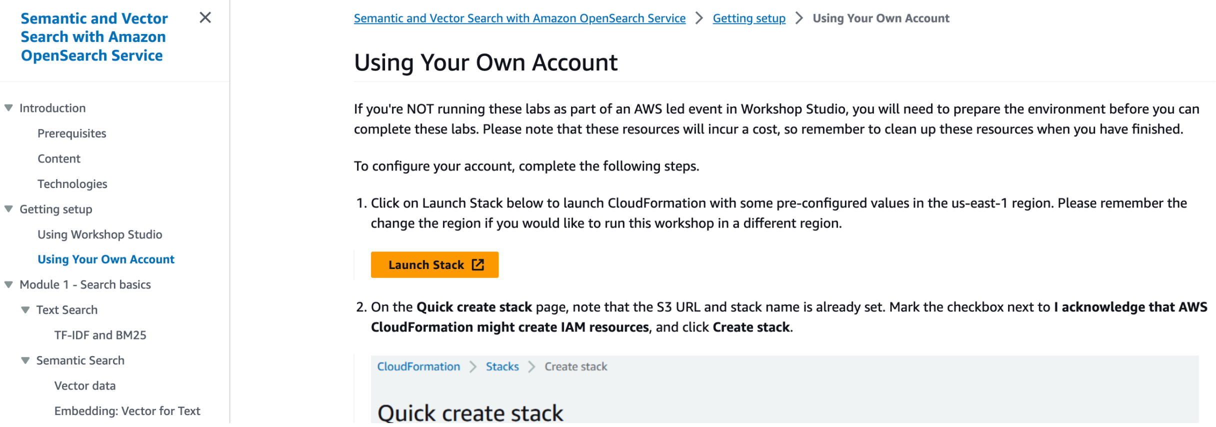
Note: the name of the end point will be different in your case



	Name	ARN	Creation time	Status	Last updated
●	opensearch-gen-ai-llm-falcon-7b-bf16-7f287940	arn:aws:sagemaker:us-east-1:057006162865:endpoint/opensearch-gen-ai-llm-falcon-7b-bf16-7f287940	9/25/2023, 7:09:25 PM	InService	9/25/2023, 7:17:24 PM
○	opensearch-gen-ai-embedding-gpt-j-6b-7f287940	arn:aws:sagemaker:us-east-1:057006162865:endpoint/opensearch-gen-ai-embedding-gpt-j-6b-7f287940	9/25/2023, 7:09:23 PM	InService	9/25/2023, 7:14:58 PM

    b. Download the dataset required for running the lab. The dataset is available in https://github.com/thandavm/rag_sm_js/blob/main/data/winemag-data-130k-v2.json

2. Launch the Cloud formation stack from https://catalog.workshops.aws/semantic-search/en-US/setup/using-own-account


Semantic and Vector Search with Amazon OpenSearch Service
Introduction
Prerequisites
Content
Technologies
Getting setup
Using Workshop Studio
Using Your Own Account
Module 1 - Search basics
Text Search
TF-IDF and BM25
Semantic Search
Vector data
Embedding: Vector for Text
Semantic and Vector Search with Amazon OpenSearch Service > Getting setup > Using Your Own Account
Launch Stack
CloudFormation > Stacks > Create stack
Quick create stack

Using Your Own Account

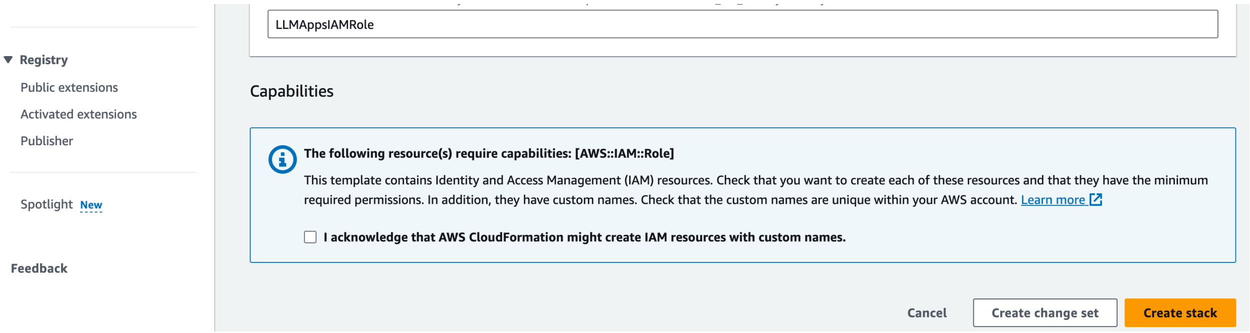
If you're NOT running these labs as part of an AWS led event in Workshop Studio, you will need to prepare the environment before you can complete these labs. Please note that these resources will incur a cost, so remember to clean up these resources when you have finished.

To configure your account, complete the following steps.

1. Click on Launch Stack below to launch CloudFormation with some pre-configured values in the us-east-1 region. Please remember the change the region if you would like to run this workshop in a different region.

2. On the **Quick create stack** page, note that the S3 URL and stack name is already set. Mark the checkbox next to **I acknowledge that AWS CloudFormation might create IAM resources**, and click **Create stack**.

3. Select "I Acknowledge…." And click on "Create Stack". The stack creation takes ~15 mins to complete



LLMAppsIAMRole

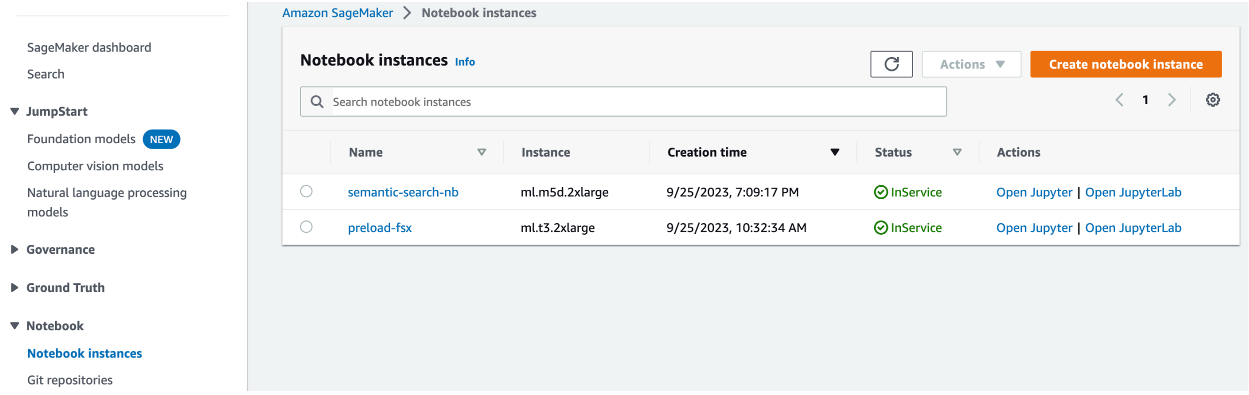
Capabilities

The following resource(s) require capabilities: [AWS::IAM::Role]

This template contains Identity and Access Management (IAM) resources. Check that you want to create each of these resources and that they have the minimum required permissions. In addition, they have custom names. Check that the custom names are unique within your AWS account. Learn more

□ **I acknowledge that AWS CloudFormation might create IAM resources with custom names.**

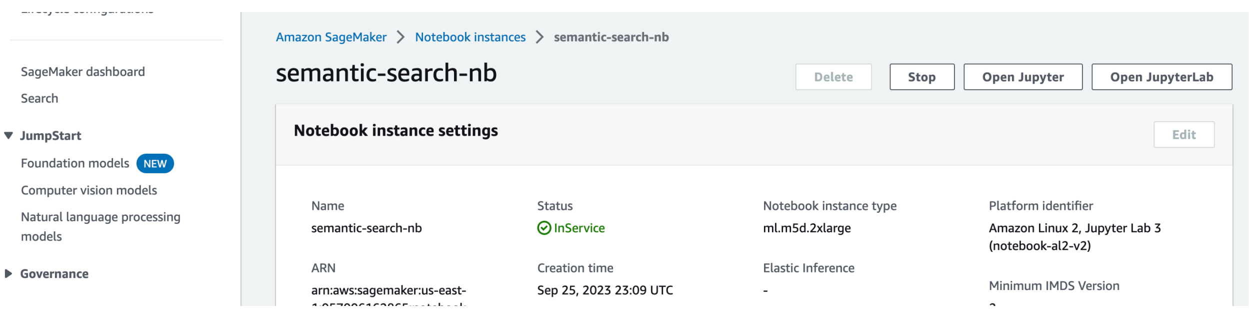
4. Once the stack is created. Go to SageMaker -> Notebook Instances.



Notebook instances

	Name	Instance	Creation time	Status	Actions
○	semantic-search-nb	ml.m5d.2xlarge	9/25/2023, 7:09:17 PM	✓ InService	Open Jupyter | Open JupyterLab
○	preload-fsx	ml.t3.2xlarge	9/25/2023, 10:32:34 AM	✓ InService	Open Jupyter | Open JupyterLab

5. Open "semantic-search-nb" and launch "Open Jupyter"

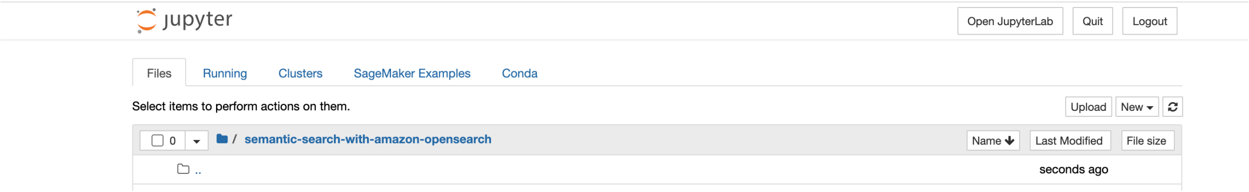


semantic-search-nb

Notebook instance settings

Name semantic-search-nb	Status ✓ InService	Notebook instance type ml.m5d.2xlarge	Platform identifier Amazon Linux 2, Jupyter Lab 3 (notebook-al2-v2)
ARN arn:aws:sagemaker:us-east-	Creation time Sep 25, 2023 23:09 UTC	Elastic Inference -	Minimum IMDS Version

6. Click on "Upload" and load the downloaded "winemag-data-130k-v2.json"

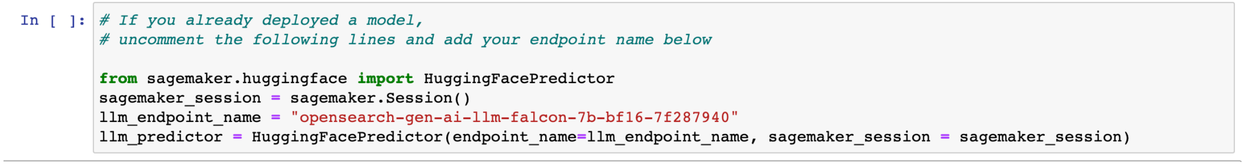


Select items to perform actions on them.

/ semantic-search-with-amazon-opensearch

.. seconds ago

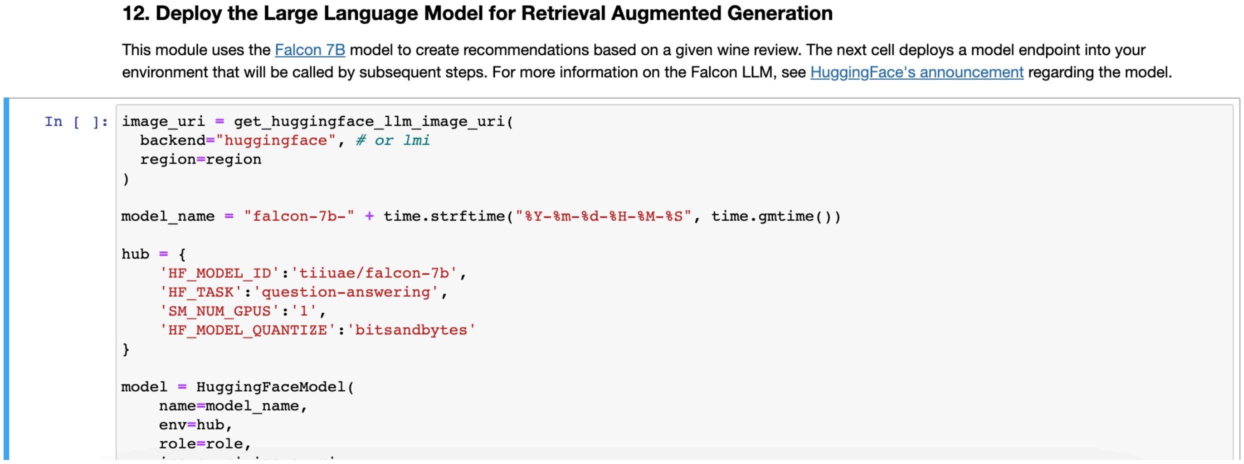
7. Open the Notebook "Module 7 - Retrieval Augmented Generation.ipynb" and add the endpoints created as part of the cloud formation template. Search for the cell below and add the SM falcon end point name here



```
In [ ]: # If you already deployed a model,
        # uncomment the following lines and add your endpoint name below

        from sagemaker.huggingface import HuggingFacePredictor
        sagemaker_session = sagemaker.Session()
        llm_endpoint_name = "opensearch-gen-ai-llm-falcon-7b-bf16-7f287940"
        llm_predictor = HuggingFacePredictor(endpoint_name=llm_endpoint_name, sagemaker_session = sagemaker_session)
```

8. Start executing the Notebook
9. Do not execute Step 12, because we have already deployed the model



12. Deploy the Large Language Model for Retrieval Augmented Generation

This module uses the Falcon 7B model to create recommendations based on a given wine review. The next cell deploys a model endpoint into your environment that will be called by subsequent steps. For more information on the Falcon LLM, see HuggingFace's announcement regarding the model.

```
In [ ]: image_uri = get_huggingface_llm_image_uri(
          backend="huggingface", # or lmi
          region=region
        )

        model_name = "falcon-7b-" + time.strftime("%Y-%m-%d-%H-%M-%S", time.gmtime())

        hub = {
            'HF_MODEL_ID':'tiiuae/falcon-7b',
            'HF_TASK':'question-answering',
            'SM_NUM_GPUS':'1',
            'HF_MODEL_QUANTIZE':'bitsandbytes'
        }

        model = HuggingFaceModel(
            name=model_name,
            env=hub,
            role=role,
```

10. Continue and complete the lab!!!