Modular RAGs

Finn Vilsbæk

fvil@eadania.dk

Subjects for today

- Azure AI Services setup
- Running the example code base

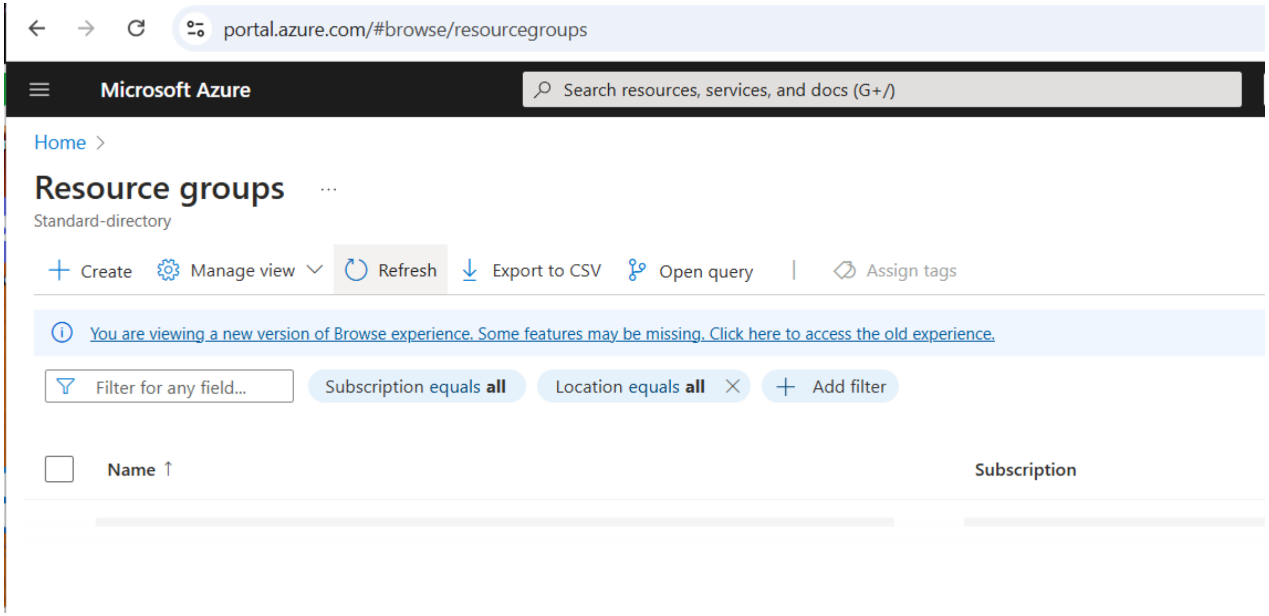
- Prerequisites: An active Azure account, preferably with a pay-as-you-go subscription and your own dedicated ressource group.

- Get the files you will need for this workshop from here: http://panmedia.dk/en-US/rag-workshop

  ..and unzip them into a local folder on your disk.

Modular RAGs

- In order to get the example code up and running on your local machine, you first need to have an active Azure Account – your student account with your institution should work fine. Create a resource group first, as you need to have a place to store your services in. You can create this by searching for ‘resource group’ in the search bar.

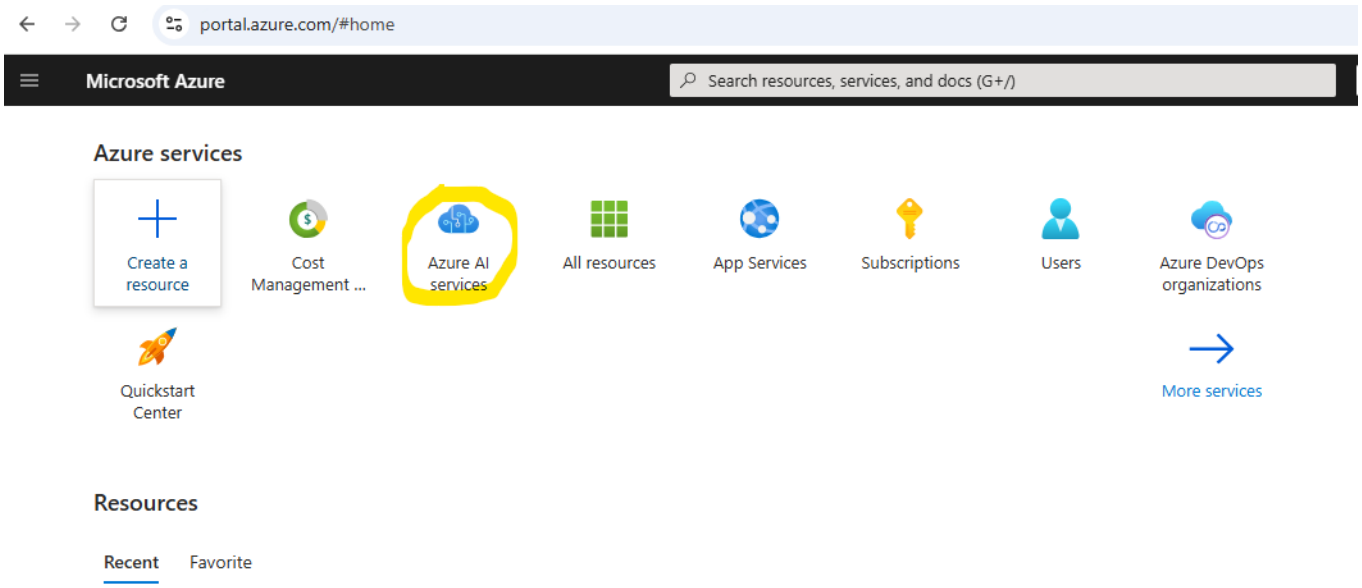


portal.azure.com/#browse/resourcegroups

You are viewing a new version of Browse experience. Some features may be missing. Click here to access the old experience.

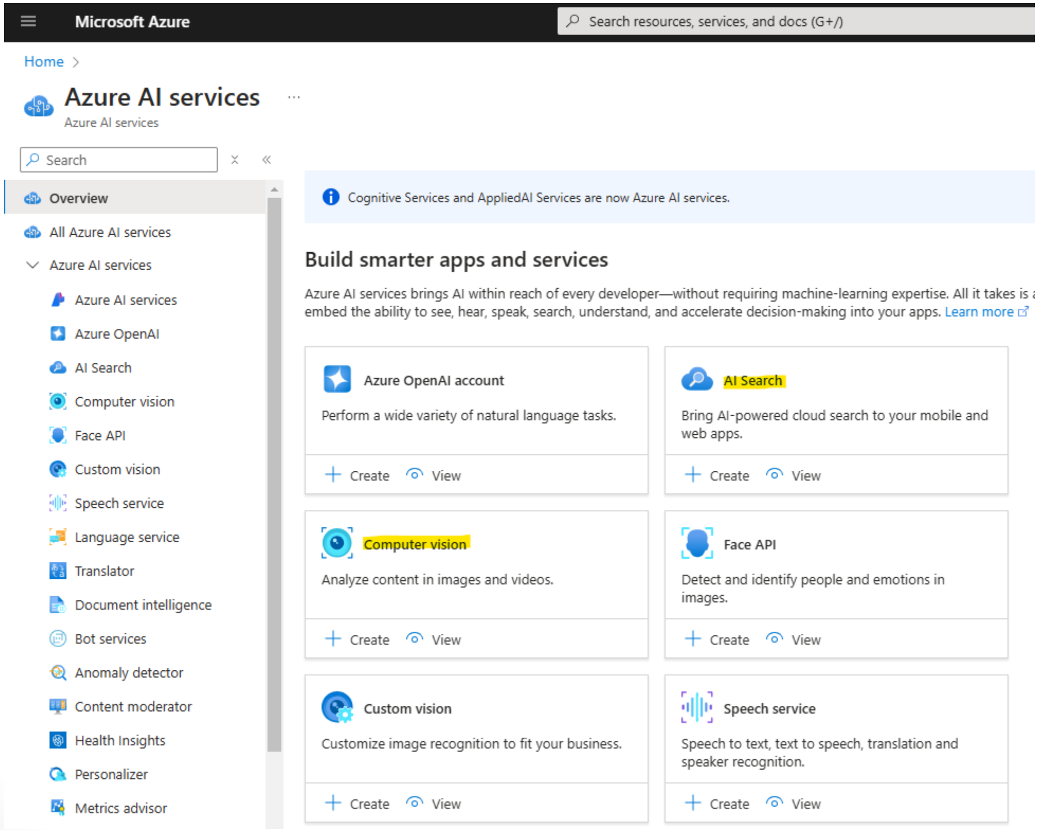
Modular RAGs

- From your Azure Portal Home Page, pick Azure AI Services. Link: https://portal.azure.com/#home



Modular RAGs

- Bob is your uncle today, because AI Search and Computer Vision, the two main services we need are pretty much at the top of the list.



Microsoft Azure

Search resources, services, and docs (G+/)

Home >

Azure AI services

Azure AI services

Search

- Overview
- All Azure AI services
- Azure AI services
  - Azure AI services
  - Azure OpenAI
  - AI Search
  - Computer vision
  - Face API
  - Custom vision
  - Speech service
  - Language service
  - Translator
  - Document intelligence
  - Bot services
  - Anomaly detector
  - Content moderator
  - Health Insights
  - Personalizer
  - Metrics advisor

Cognitive Services and AppliedAI Services are now Azure AI services.

Build smarter apps and services

Azure AI services brings AI within reach of every developer—without requiring machine-learning expertise. All it takes is a
embed the ability to see, hear, speak, search, understand, and accelerate decision-making into your apps. Learn more

Azure OpenAI account
Perform a wide variety of natural language tasks.
+ Create View

AI Search
Bring AI-powered cloud search to your mobile and web apps.
+ Create View

Computer vision
Analyze content in images and videos.
+ Create View

Face API
Detect and identify people and emotions in images.
+ Create View

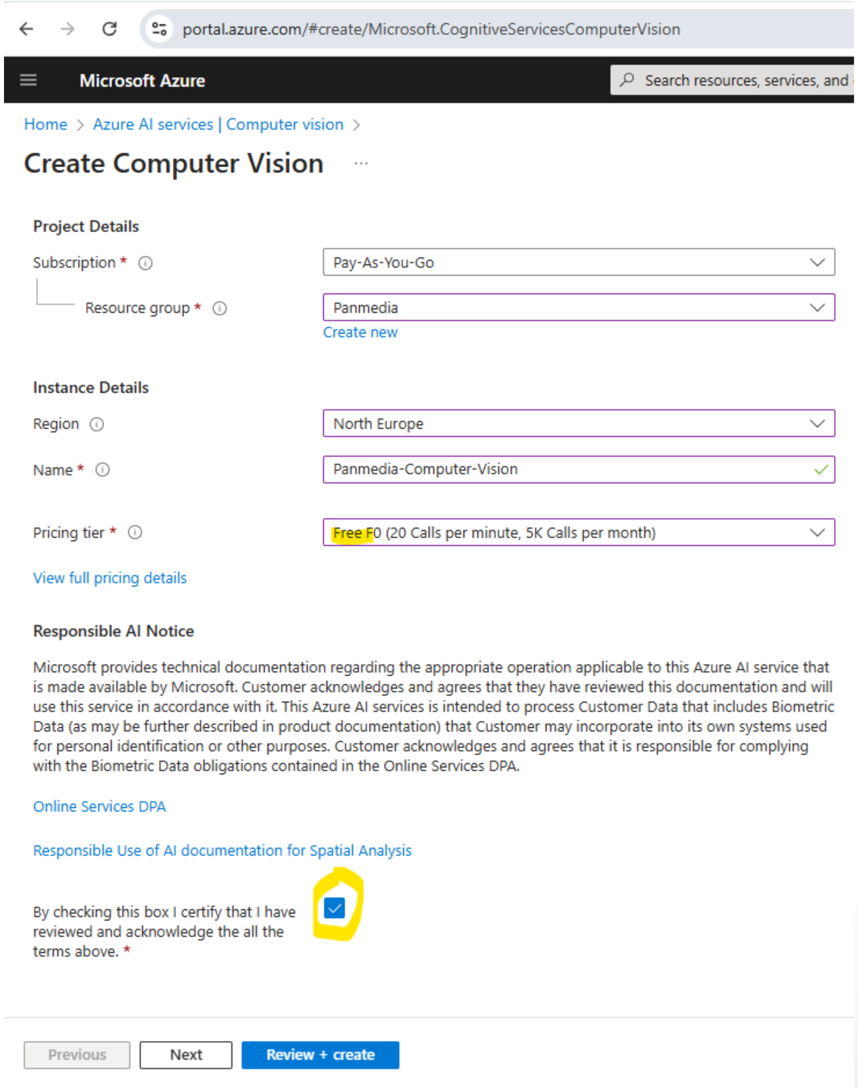
Custom vision
Customize image recognition to fit your business.
+ Create View

Speech service
Speech to text, text to speech, translation and speaker recognition.
+ Create View

Modular RAGs

- Create a Computer Vision instance.

  Be sure to pick the free F0 tier.


portal.azure.com/#create/Microsoft.CognitiveServicesComputerVision
Microsoft Azure
Search resources, services, and
Home > Azure AI services | Computer vision >

Create Computer Vision ...

Project Details

Subscription * : Pay-As-You-Go

Resource group * : Panmedia

Instance Details

Region : North Europe

Name * : Panmedia-Computer-Vision ✓

Pricing tier * : Free F0 (20 Calls per minute, 5K Calls per month)

View full pricing details

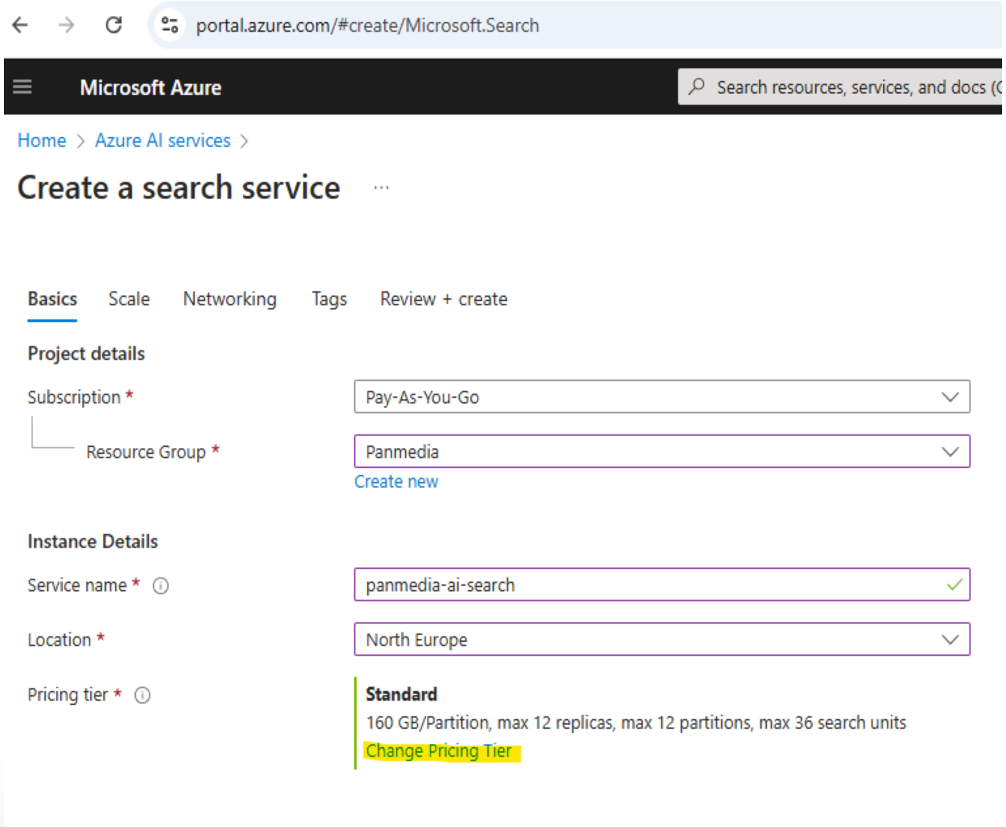
Responsible AI Notice

Microsoft provides technical documentation regarding the appropriate operation applicable to this Azure AI service that is made available by Microsoft. Customer acknowledges and agrees that they have reviewed this documentation and will use this service in accordance with it. This Azure AI services is intended to process Customer Data that includes Biometric Data (as may be further described in product documentation) that Customer may incorporate into its own systems used for personal identification or other purposes. Customer acknowledges and agrees that it is responsible for complying with the Biometric Data obligations contained in the Online Services DPA.

By checking this box I certify that I have reviewed and acknowledge the all the terms above. * ☑

Modular RAGs

- The Azure AI Search service is a little trickier to set up. Change the price tier via the link at the bottom of the page from standard to..



portal.azure.com/#create/Microsoft.Search

Home > Azure AI services >

Create a search service

Project details

Subscription * Pay-As-You-Go

Resource Group * Panmedia

Create new

Instance Details

Service name * panmedia-ai-search

Location * North Europe

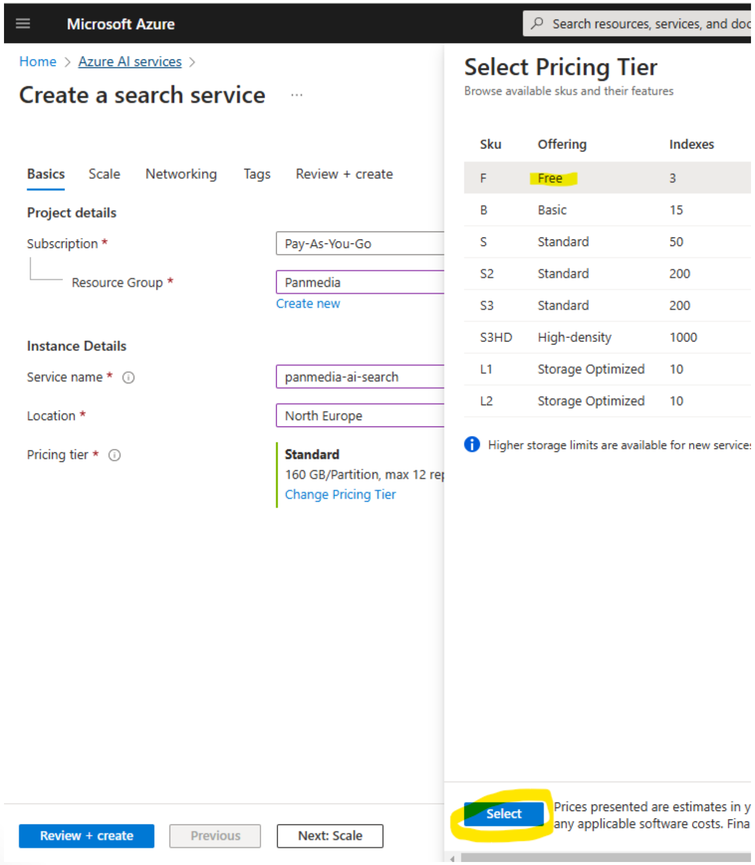
Pricing tier * **Standard**

160 GB/Partition, max 12 replicas, max 12 partitions, max 36 search units

Change Pricing Tier

Modular RAGs

- .. free, which offers you a whopping three indexes – more than we need!



Microsoft Azure
Search resources, services, and doc
Home > Azure AI services >
Basics
Scale
Networking
Tags
Review + create

Create a search service ...

Project details

Subscription * Pay-As-You-Go

Resource Group * Panmedia

Create new

Instance Details

Service name * panmedia-ai-search

Location * North Europe

Pricing tier * **Standard**
160 GB/Partition, max 12 rep
Change Pricing Tier

Select Pricing Tier

Browse available skus and their features

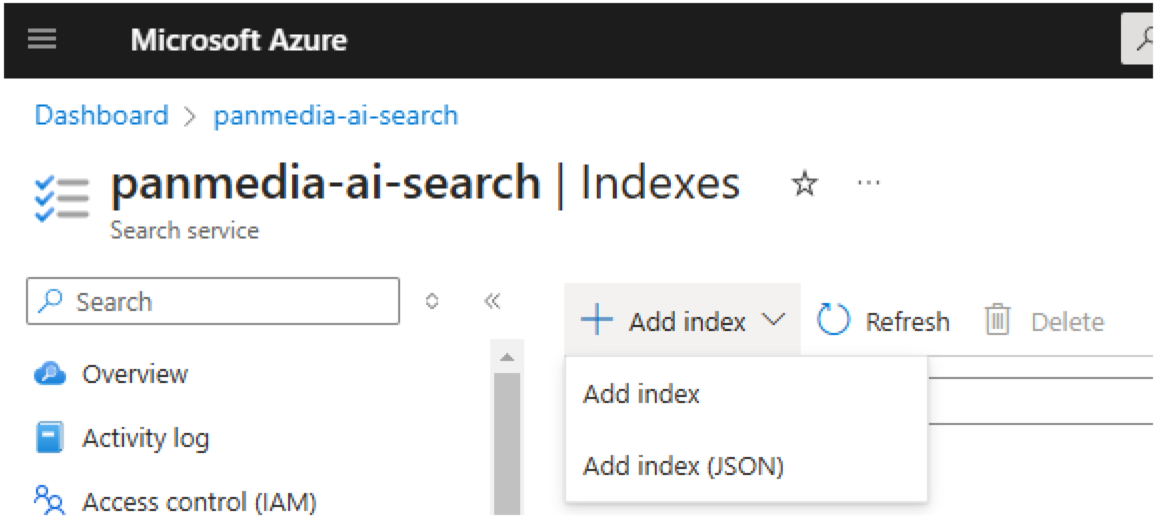
Sku	Offering	Indexes
F	Free	3
B	Basic	15
S	Standard	50
S2	Standard	200
S3	Standard	200
S3HD	High-density	1000
L1	Storage Optimized	10
L2	Storage Optimized	10

Higher storage limits are available for new service

Prices presented are estimates in y
any applicable software costs. Fina

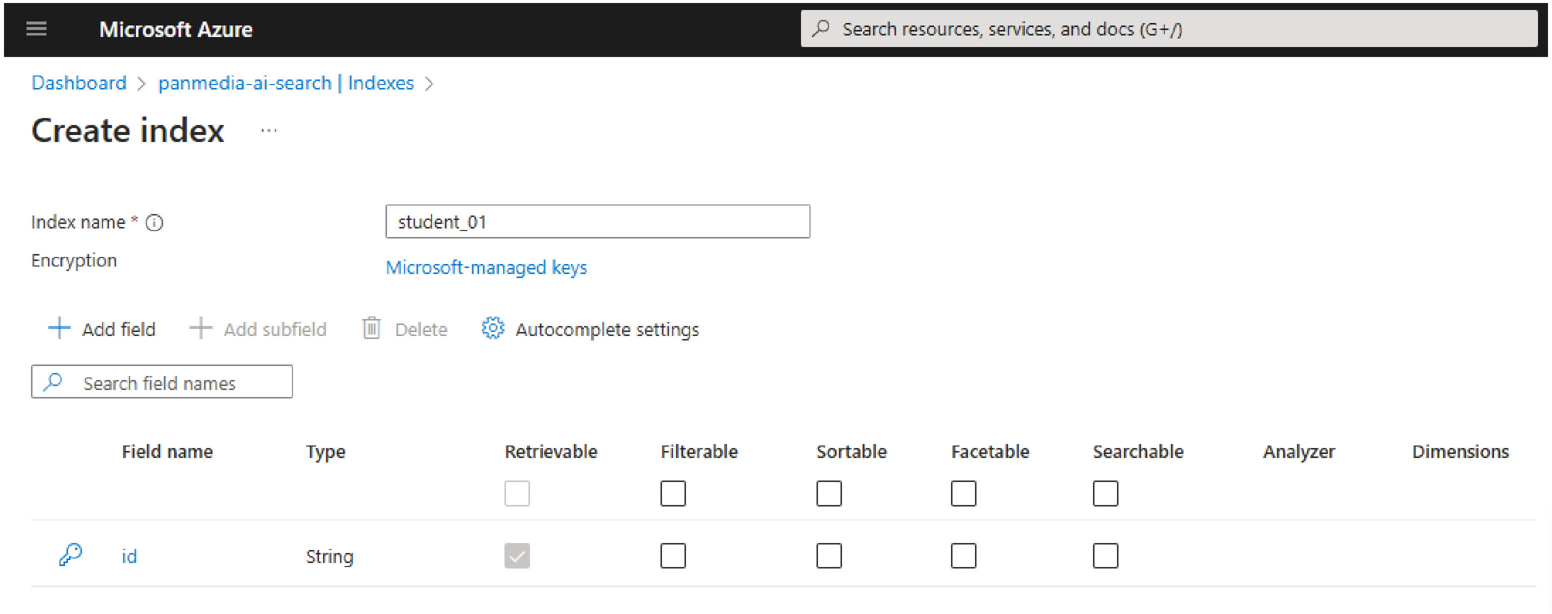
Modular RAGs

- Setting up a search index is your next task. Simply go to your AI Search instance when the resource is created, and click on ‘Add Index’ – choose the normal index, not the JSON variant.



Modular RAGs

- Give your Index a memorable name, and click on ‘Add Field’.



Microsoft Azure
Search resources, services, and docs (G+/)
Dashboard > panmedia-ai-search | Indexes >

Create index

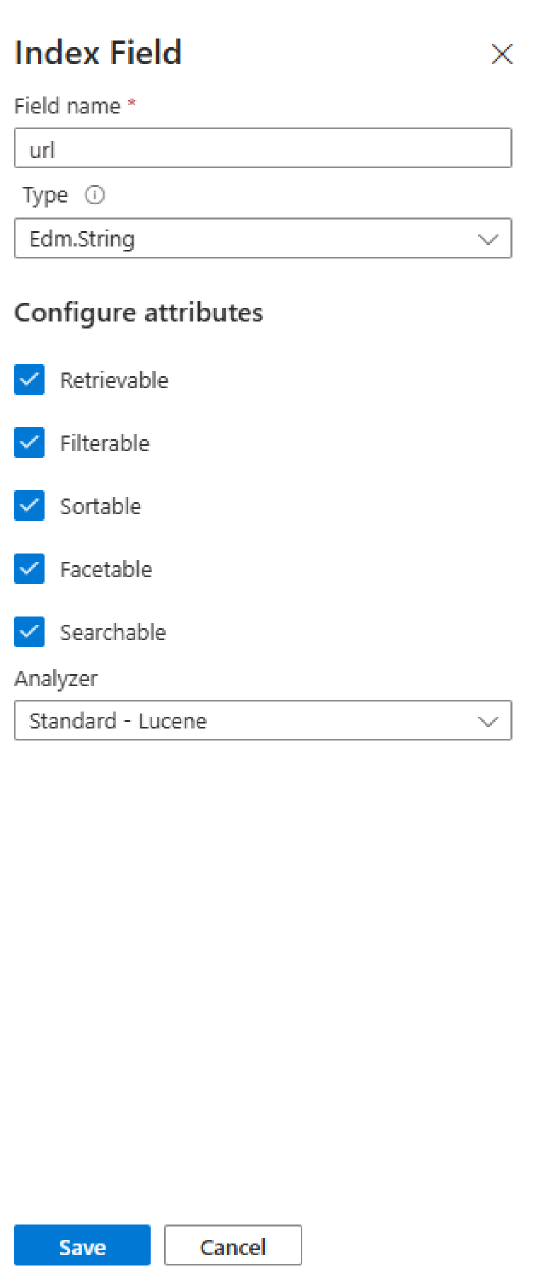
Index name * ⓘ: student_01

Encryption: Microsoft-managed keys

	Field name	Type	Retrievable	Filterable	Sortable	Facetable	Searchable	Analyzer	Dimensions
			□	□	□	□	□		
🔑	id	String	☑	□	□	□	□		

Modular RAGs

- Create an url field. Check all the boxes, so the field gets a standard Lucene analyzer.



Index Field

Field name *

url

Type

Edm.String

Configure attributes

☑ Retrievable

☑ Filterable

☑ Sortable

☑ Facetable

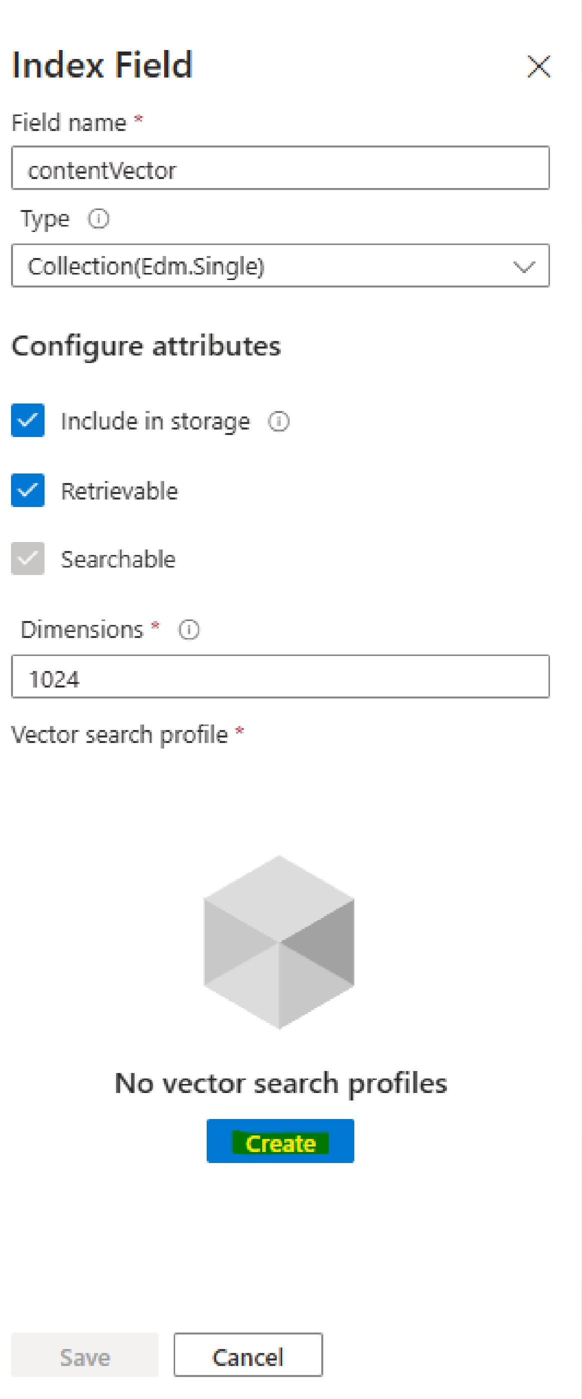
☑ Searchable

Analyzer

Standard - Lucene

Modular RAGs

- Create a field ‘contentVector’ of type Collection(Edm.Single) with the settings shown here, and also create a vector search profile.



Index Field

Field name *

contentVector

Type

Collection(Edm.Single)

Configure attributes

☑ Include in storage

☑ Retrievable

☑ Searchable

Dimensions *

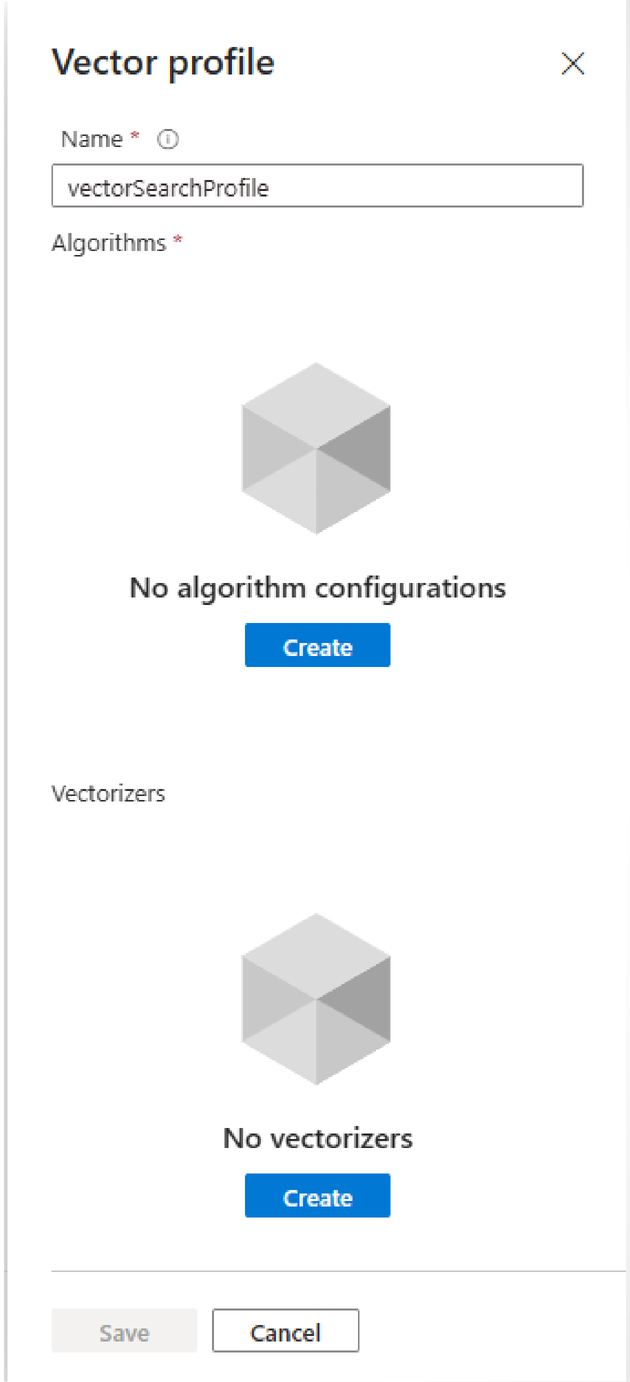
1024

Vector search profile *

No vector search profiles

Modular RAGs

- You will also need to create an algorithm configuration for your Vector Search Profile.



Vector profile

Name * ⓘ

vectorSearchProfile

Algorithms *

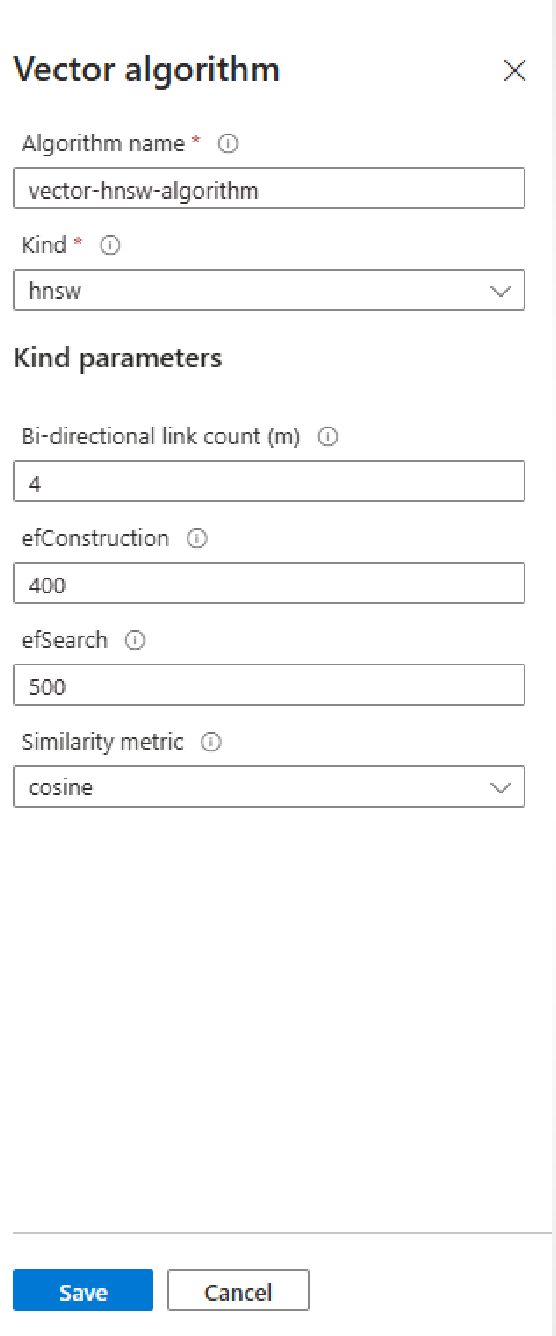
No algorithm configurations

Vectorizers

No vectorizers

Modular RAGs

- The Vector algorithm should have the settings shown here:



Vector algorithm

Algorithm name *

vector-hnsw-algorithm

Kind *

hnsw

Kind parameters

Bi-directional link count (m)

4

efConstruction

400

efSearch

500

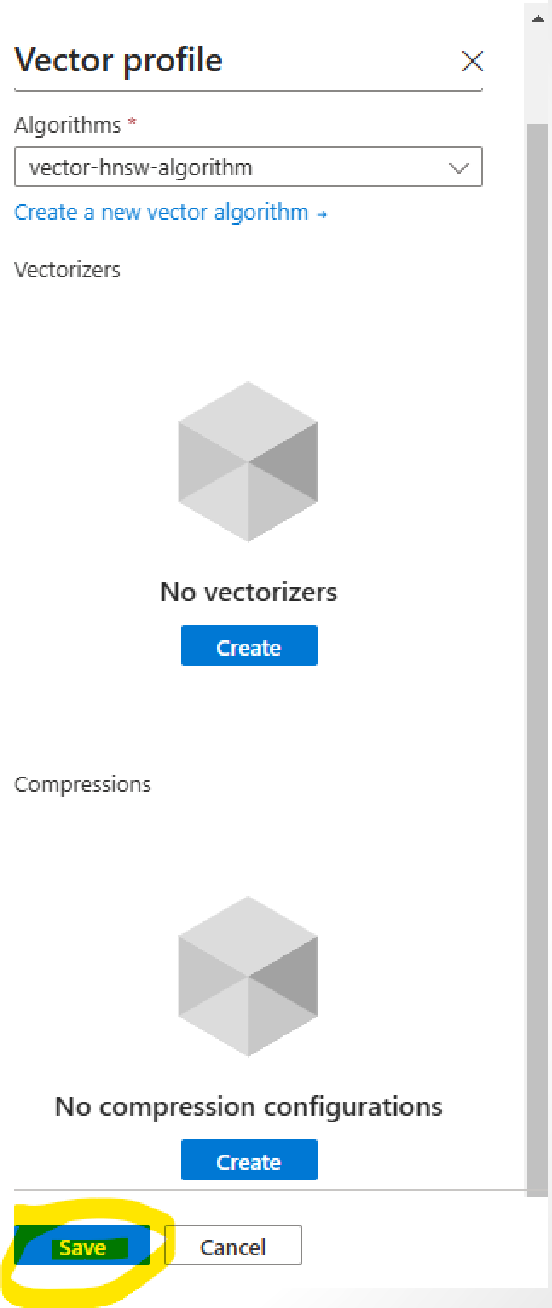
Similarity metric

cosine

Modular RAGs

- Now you need to save your new Vector profile – make sure that your algorithm is selected in the dropdown box and click Save.

  You don't need to create a Vectorizer or a Compression configuration.



Vector profile

Algorithms *

vector-hnsw-algorithm

Vectorizers

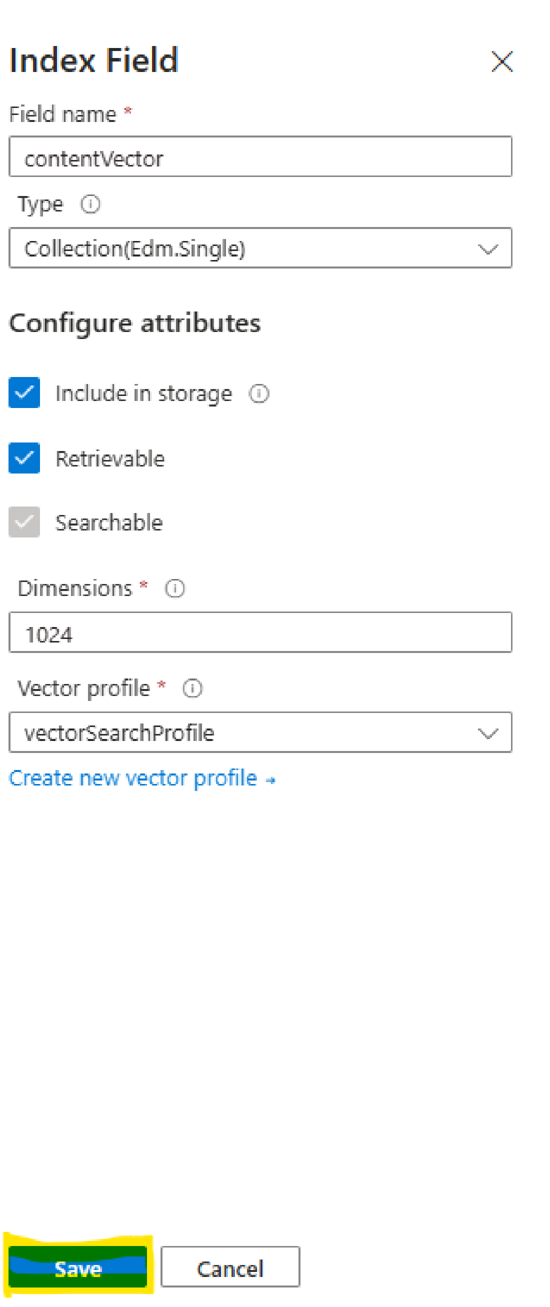
No vectorizers

Compressions

No compression configurations

Modular RAGs

- Now the system will return to the previous screen, and you can save the contentVector index field, along with the Vector profile you have just created. NB: **make sure** that the boxes are ticked here!



Index Field

Field name *

contentVector

Type

Collection(Edm.Single)

Configure attributes

☑ Include in storage

☑ Retrievable

☑ Searchable

Dimensions *

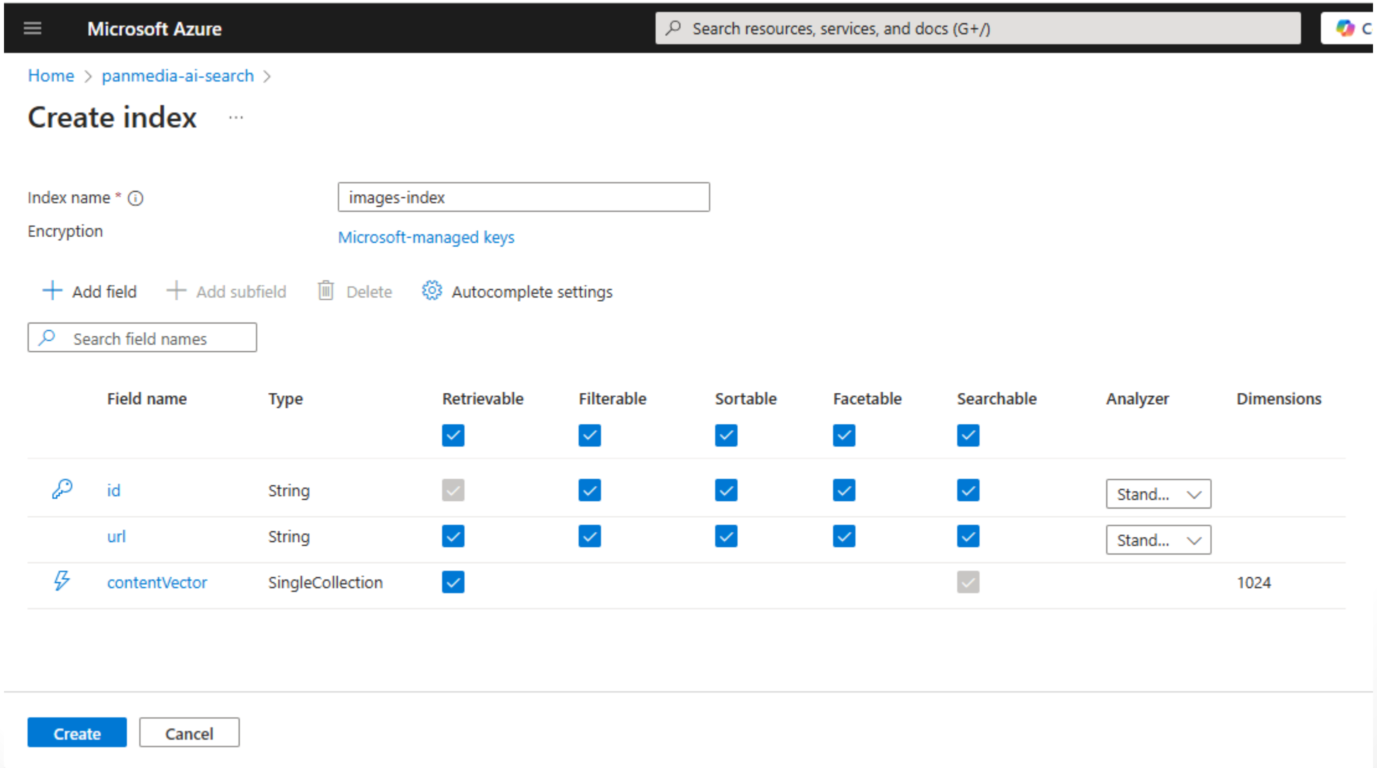
1024

Vector profile *

vectorSearchProfile

Modular RAGs

- Lastly, you need to create the index. If you have made the index precisely as indicated in the previous screendumps, the boxes are ticked as shown here, and you can hit ‘Create’. **NB: if you click any of the checkboxes in this view, you have to make the index from the starting point again** – it’s a bug in the Azure interface, sorry guys ☹



Microsoft Azure
Search resources, services, and docs (G+/)
Home > panmedia-ai-search >

Create index ···

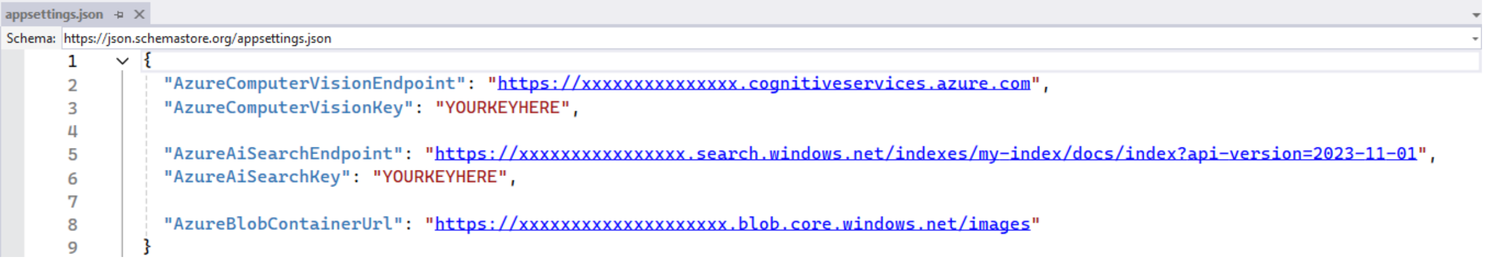
Index name * ⓘ images-index

Encryption Microsoft-managed keys

Field name	Type	Retrievable	Filterable	Sortable	Facetable	Searchable	Analyzer	Dimensions
		☑	☑	☑	☑	☑		
id	String	☑	☑	☑	☑	☑	Stand...	
url	String	☑	☑	☑	☑	☑	Stand...	
contentVector	SingleCollection	☑				☑		1024

Modular RAGs

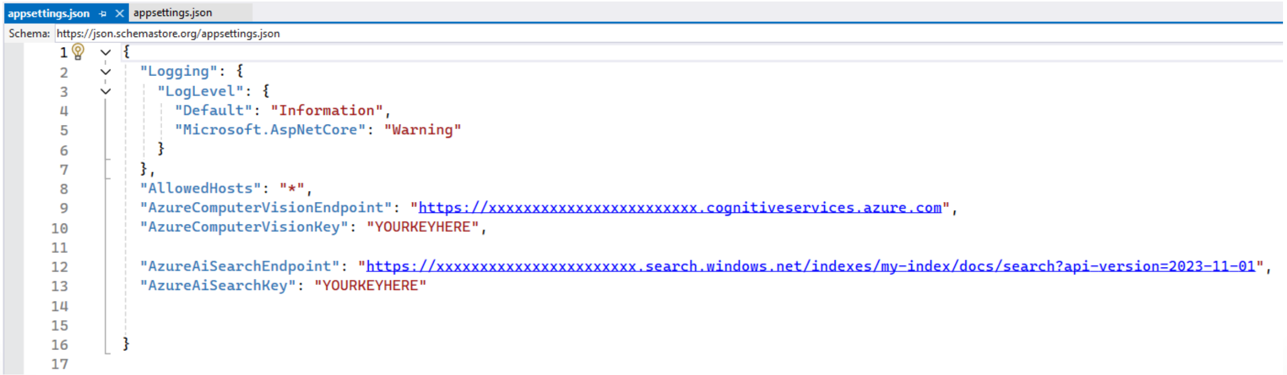
- Now, you can set up your unique API keys and endpoints from your own Azure account in the two solution projects. The appsettings.json file in the Console project:



Schema: https://json.schemastore.org/appsettings.json

```
{
  "AzureComputerVisionEndpoint": "https://xxxxxxxxxxxxxxx.cognitiveservices.azure.com",
  "AzureComputerVisionKey": "YOURKEYHERE",

  "AzureAiSearchEndpoint": "https://xxxxxxxxxxxxxxxx.search.windows.net/indexes/my-index/docs/index?api-version=2023-11-01",
  "AzureAiSearchKey": "YOURKEYHERE",

  "AzureBlobContainerUrl": "https://xxxxxxxxxxxxxxxxxxxx.blob.core.windows.net/images"
}
```

.. and the appsettings.json file in the WebApplication project:

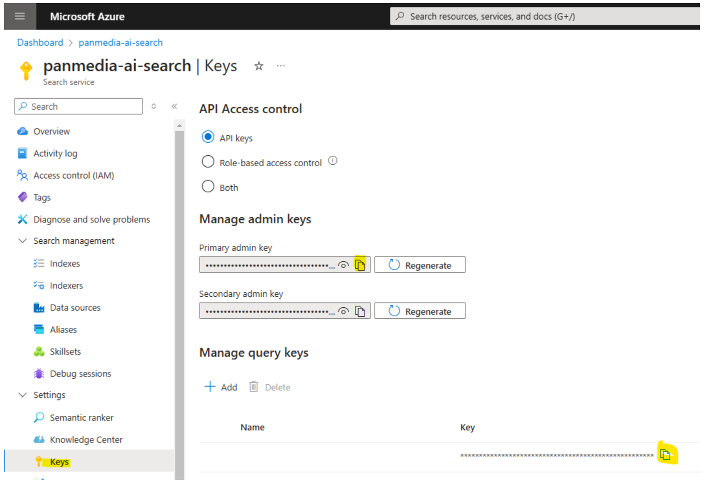


Schema: https://json.schemastore.org/appsettings.json

```
{
  "Logging": {
    "LogLevel": {
      "Default": "Information",
      "Microsoft.AspNetCore": "Warning"
    }
  },
  "AllowedHosts": "*",
  "AzureComputerVisionEndpoint": "https://xxxxxxxxxxxxxxxxxxxxxxxxx.cognitiveservices.azure.com",
  "AzureComputerVisionKey": "YOURKEYHERE",

  "AzureAiSearchEndpoint": "https://xxxxxxxxxxxxxxxxxxxxxxxx.search.windows.net/indexes/my-index/docs/search?api-version=2023-11-01",
  "AzureAiSearchKey": "YOURKEYHERE"



}
```

Modular RAGs

You can copy the keys and endpoints you need from each resource from Settings >> Keys and Resource Management >> Keys and Endpoints. Click on the icon marked with yellow to copy the key to the clipboard.



panmedia-ai-search | Keys

Search service

API Access control

● API keys

○ Role-based access control

○ Both

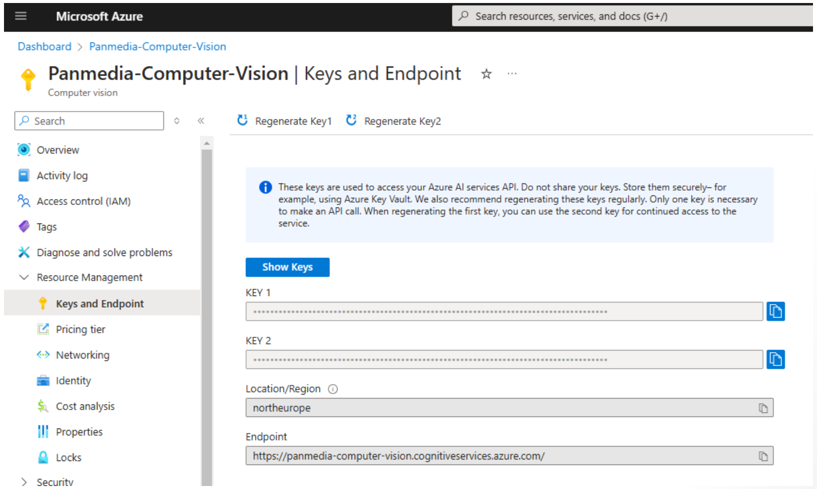
Manage admin keys

Primary admin key: ••••••••••••••••••••••••••••••••...

Secondary admin key: ••••••••••••••••••••••••••••••••...

Manage query keys

Name	Key
	••



Panmedia-Computer-Vision | Keys and Endpoint

Computer vision

These keys are used to access your Azure AI services API. Do not share your keys. Store them securely– for example, using Azure Key Vault. We also recommend regenerating these keys regularly. Only one key is necessary to make an API call. When regenerating the first key, you can use the second key for continued access to the service.

KEY 1: ••

KEY 2: ••

Location/Region: northeurope

Endpoint: https://panmedia-computer-vision.cognitiveservices.azure.com/

Modular RAGs

NB: the search endpoint in the console app indexer is special:

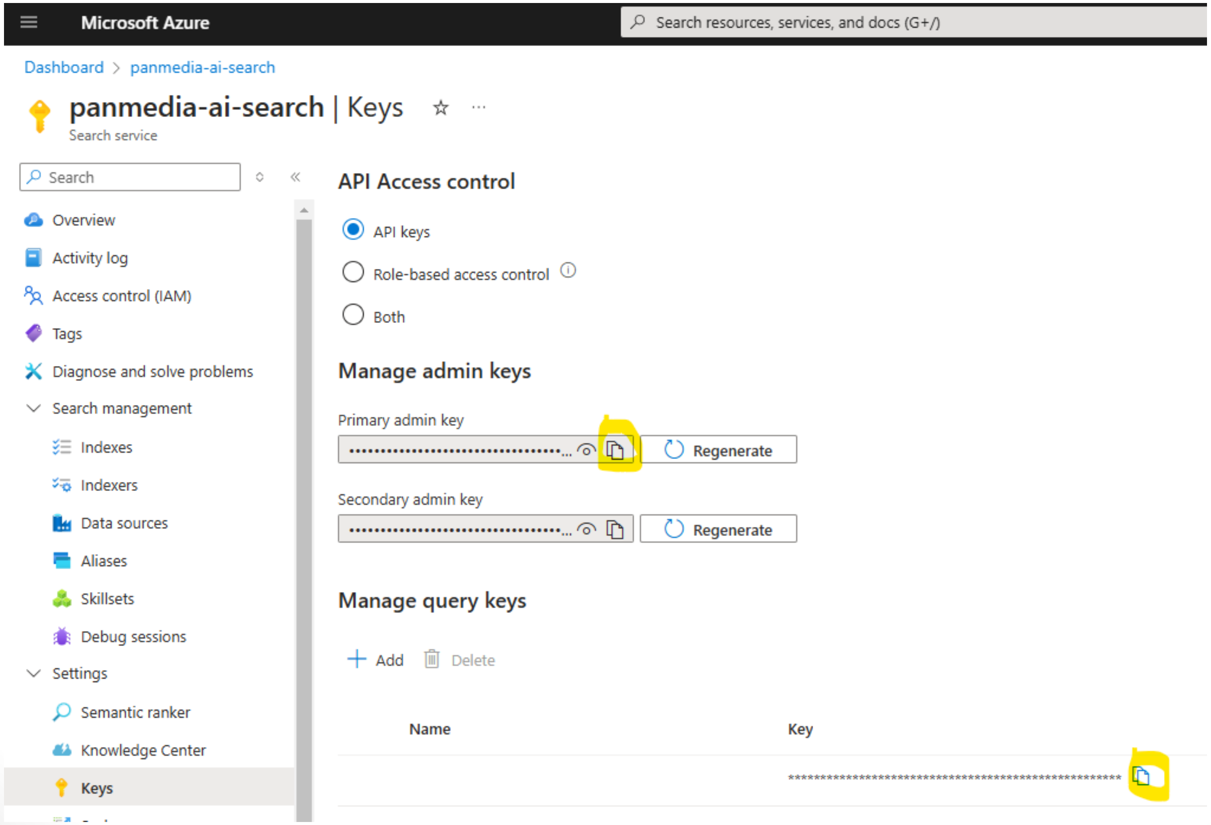
"AzureAiSearchEndpoint":

https://YOUR_NAMED_SEARCH_INSTANCE.search.windows.net/indexes/YOUR_NAMED_INDEX/docs/index?api-version=2023-11-01

Here, you can set up your own names instead of the capital letters.

Modular RAGs

- Important: set your primary admin key for ai search in the Console project, and use the more lowly query key in the Web project. If you don't, the Console App will not index correctly.



panmedia-ai-search | Keys

Search service

API Access control

● API keys

○ Role-based access control

○ Both

Manage admin keys

Primary admin key: ••••••••••••••••••••••••••••••••••...

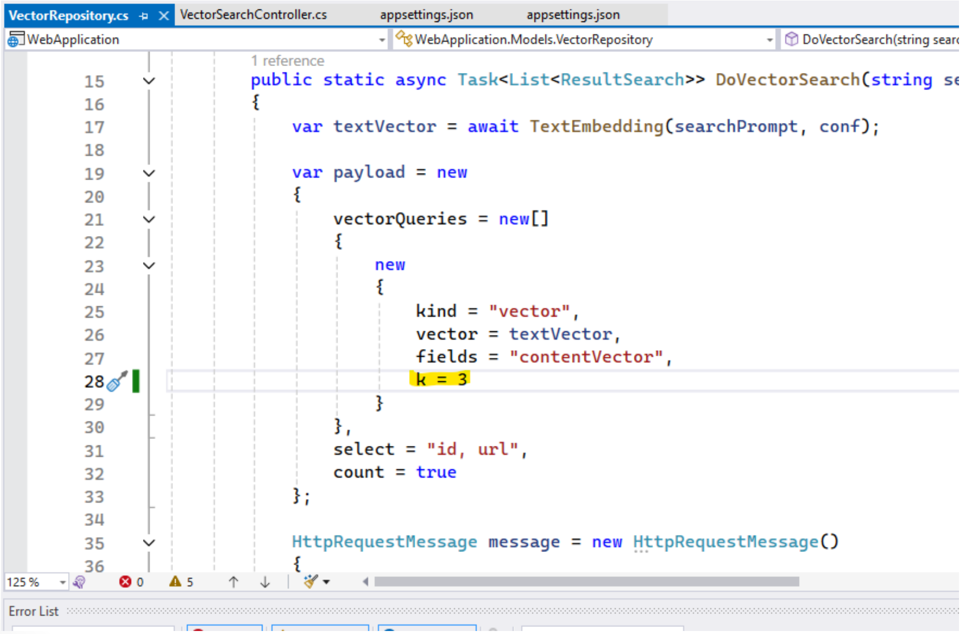
Secondary admin key: ••••••••••••••••••••••••••••••••••...

Manage query keys

Name	Key
	**

Modular RAGs

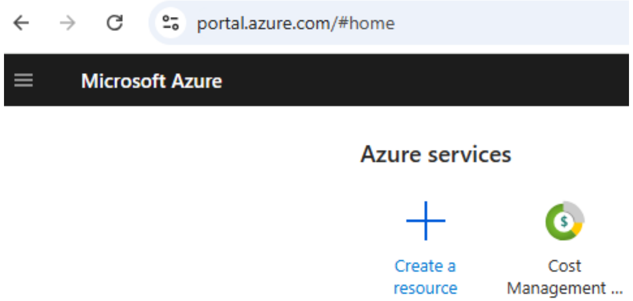
- In the file ‘VectorRepository.cs’ in the Web project, you can play around with the value for k in line 28 to 4 or 5 instead of 3. This will later bring up three, four or five pictures from the urls in the search index. There aren’t many images though in the sample folder, and with the value of 3 you will see less ‘hallucinatory’ pictures at the bottom of the list ☺

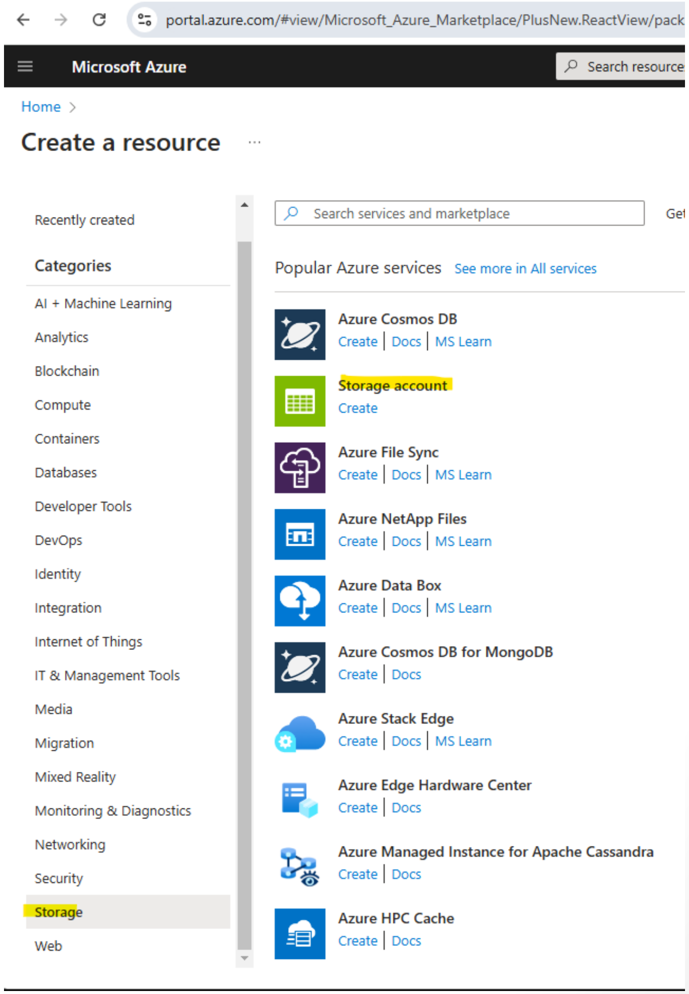

VectorRepository.cs
VectorSearchController.cs
appsettings.json
appsettings.json
WebApplication
WebApplication.Models.VectorRepository
DoVectorSearch(string searc
125 %
0
5
Error List

```
                 1 reference
15  public static async Task<List<ResultSearch>> DoVectorSearch(string se
16  {
17      var textVector = await TextEmbedding(searchPrompt, conf);
18
19      var payload = new
20      {
21          vectorQueries = new[]
22          {
23              new
24              {
25                  kind = "vector",
26                  vector = textVector,
27                  fields = "contentVector",
28                  k = 3
29              }
30          },
31          select = "id, url",
32          count = true
33      };
34
35      HttpRequestMessage message = new HttpRequestMessage()
36      {
```

Modular RAGs

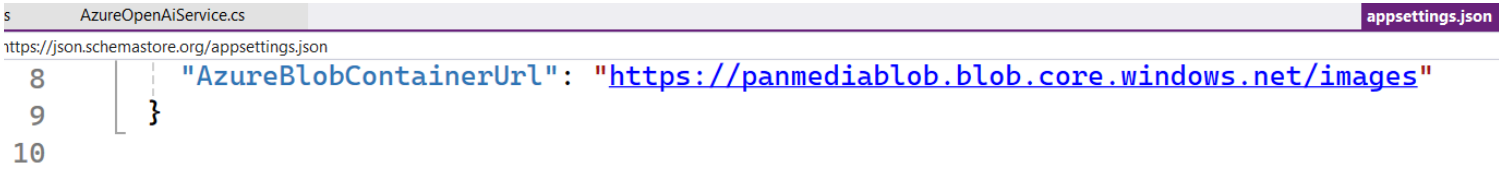
- Now for the pretty pictures part. You will need to set up a small blob storage on Azure, since the indexing we need to do is a lot easier on that platform. Go to https://portal.azure.com/#home, and click on 'Create a resource'. Then choose 'Storage Account'.





Modular RAGs

If you are unable to create a Blob storage container on your standard Azure student account – I know this can be pretty troublesome - then you can use my Blob storage account in the appsettings.json file:

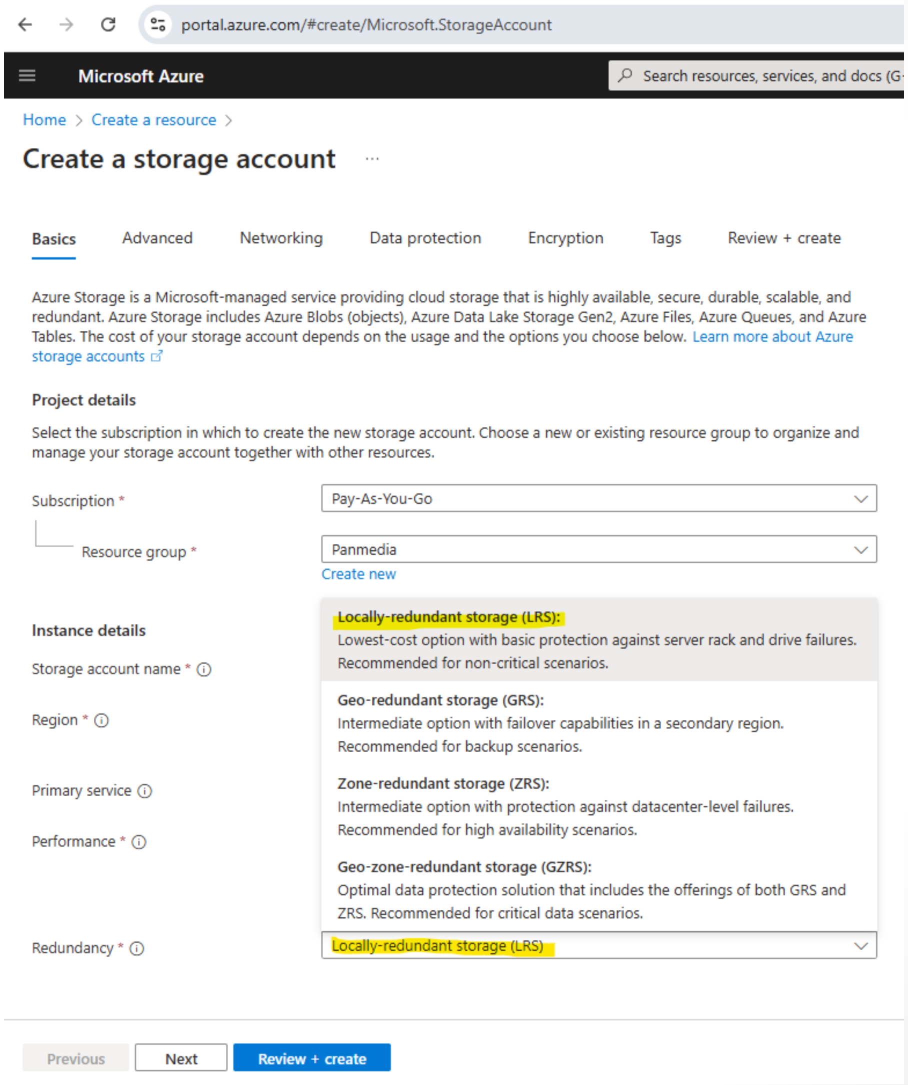


```
    "AzureBlobContainerUrl": "https://panmediablob.blob.core.windows.net/images"
}
```

If you want to try to create your own Blob storage container, you can proceed with the next six slides.

Modular RAGs

- I chose the cheapest possible options. Don't worry too much here, it most probably isn't going to break the bank account, but it is necessary in order to make the App work.



portal.azure.com/#create/Microsoft.StorageAccount

Microsoft Azure
Search resources, services, and docs (G·
Home > Create a resource >

Create a storage account ···

Azure Storage is a Microsoft-managed service providing cloud storage that is highly available, secure, durable, scalable, and redundant. Azure Storage includes Azure Blobs (objects), Azure Data Lake Storage Gen2, Azure Files, Azure Queues, and Azure Tables. The cost of your storage account depends on the usage and the options you choose below. Learn more about Azure storage accounts

Project details

Select the subscription in which to create the new storage account. Choose a new or existing resource group to organize and manage your storage account together with other resources.

Subscription *: Pay-As-You-Go

Resource group *: Panmedia

Create new

Instance details

Storage account name *

Region *

Primary service

Performance *

Locally-redundant storage (LRS):
Lowest-cost option with basic protection against server rack and drive failures. Recommended for non-critical scenarios.

Geo-redundant storage (GRS):
Intermediate option with failover capabilities in a secondary region. Recommended for backup scenarios.

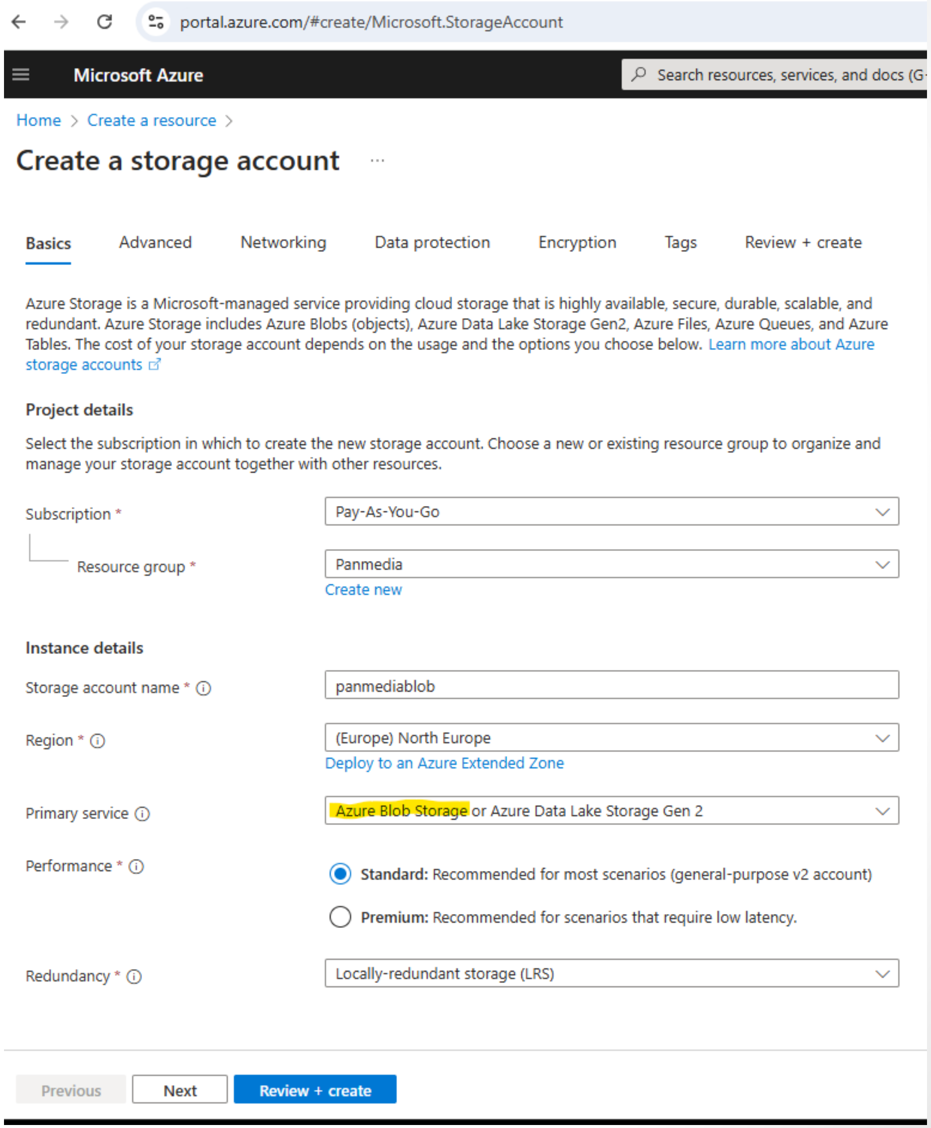
Zone-redundant storage (ZRS):
Intermediate option with protection against datacenter-level failures. Recommended for high availability scenarios.

Geo-zone-redundant storage (GZRS):
Optimal data protection solution that includes the offerings of both GRS and ZRS. Recommended for critical data scenarios.

Redundancy *: Locally-redundant storage (LRS)

Modular RAGs

- Check your settings – make sure that your storage region is relevant to the region that your other resources belong in, and make sure that your primary service is set to include 'Azure Blob Storage'.



portal.azure.com/#create/Microsoft.StorageAccount

Create a storage account

Azure Storage is a Microsoft-managed service providing cloud storage that is highly available, secure, durable, scalable, and redundant. Azure Storage includes Azure Blobs (objects), Azure Data Lake Storage Gen2, Azure Files, Azure Queues, and Azure Tables. The cost of your storage account depends on the usage and the options you choose below. Learn more about Azure storage accounts

Project details

Select the subscription in which to create the new storage account. Choose a new or existing resource group to organize and manage your storage account together with other resources.

Subscription * : Pay-As-You-Go

Resource group * : Panmedia

Create new

Instance details

Storage account name * : panmediablob

Region * : (Europe) North Europe

Deploy to an Azure Extended Zone

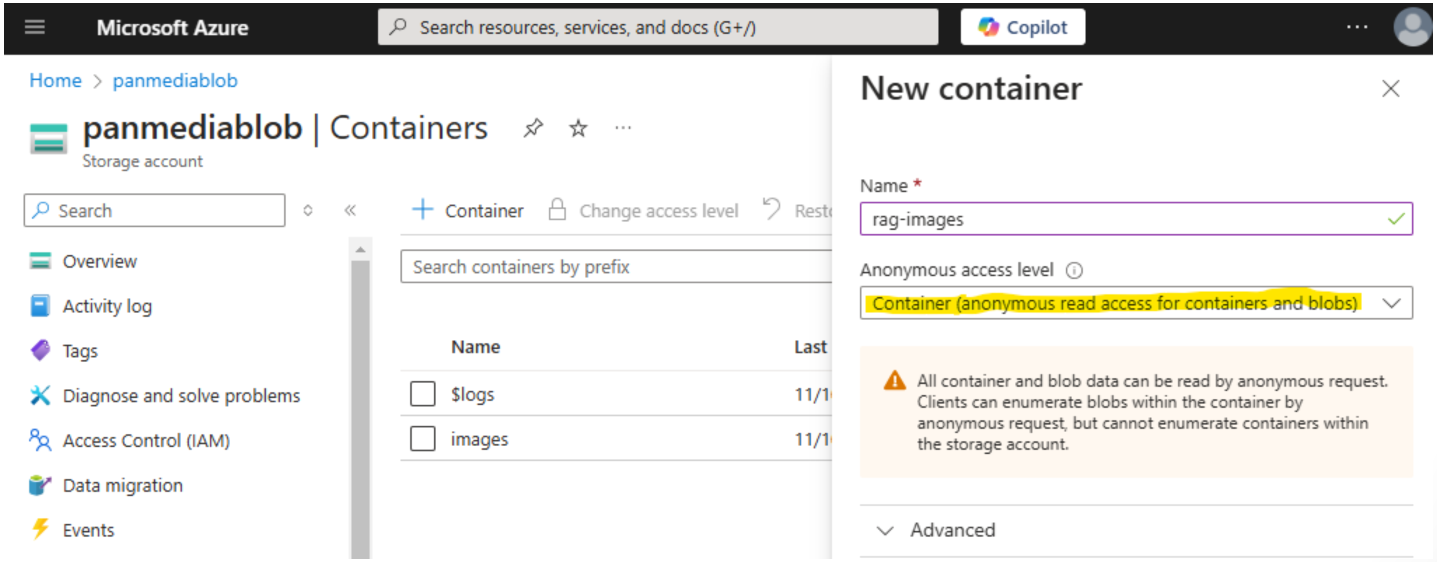
Primary service : Azure Blob Storage or Azure Data Lake Storage Gen 2

Performance * :
● **Standard:** Recommended for most scenarios (general-purpose v2 account)
○ **Premium:** Recommended for scenarios that require low latency.

Redundancy * : Locally-redundant storage (LRS)

Modular RAGs

- Next, we will need a container for our images. Click on the plus sign next to ‘Container’, and select the most permissive anonymous access level, since we want our Console App to be able to see and enumerate the container contents.



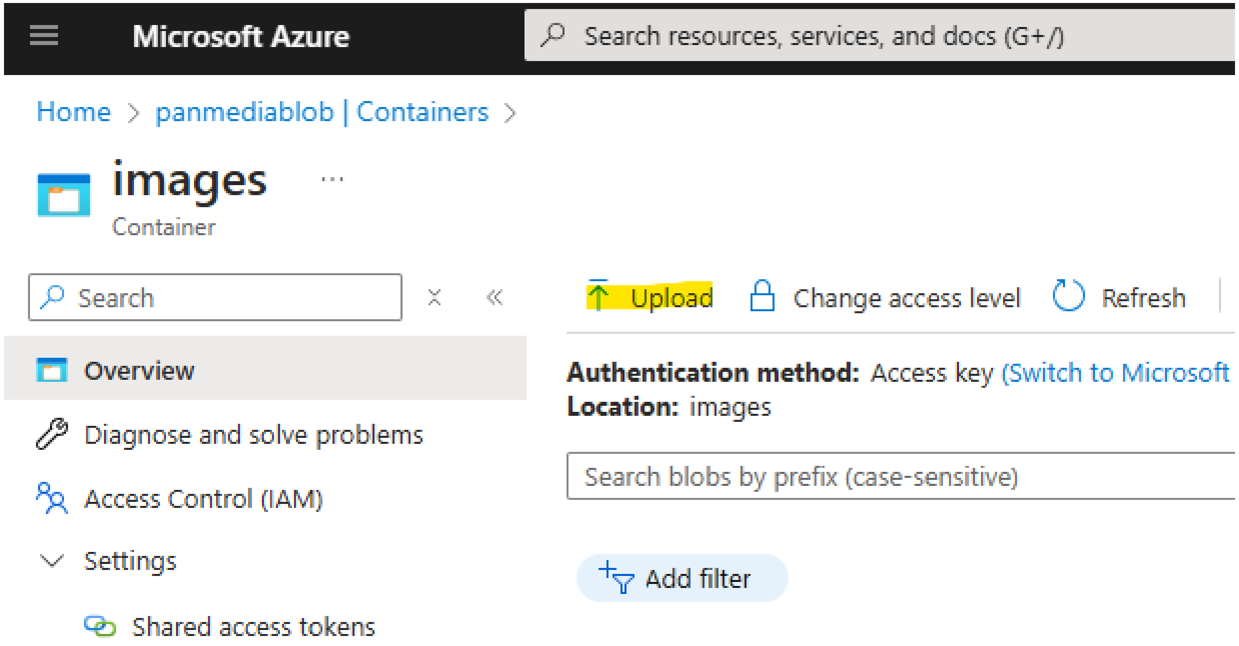
Name *: rag-images

Anonymous access level: Container (anonymous read access for containers and blobs)

All container and blob data can be read by anonymous request. Clients can enumerate blobs within the container by anonymous request, but cannot enumerate containers within the storage account.

Modular RAGs

- Take the images from the zip file you have downloaded - and upload them into your container. There should be 60 images in all.



Microsoft Azure

Search resources, services, and docs (G+/)

Home > panmediablob | Containers >

images ...

Container

Search

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Upload | Change access level | Refresh

Authentication method: Access key (Switch to Microsoft

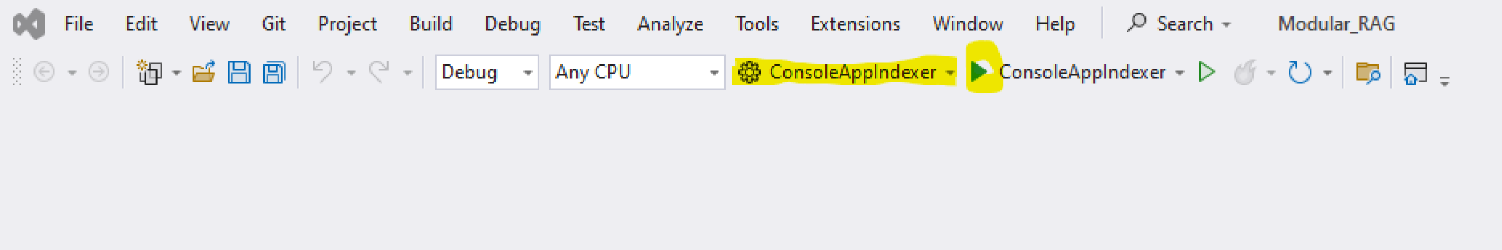
Location: images

Search blobs by prefix (case-sensitive)

Add filter

Modular RAGs

- Now, with the keys and endpoints correctly set up in both instances of appsettings.json, we should be able to run the indexer from the console. The project you have downloaded will need to be set to ’ConsoleAppIndexer’ in the dropdown in VS 2022, and you can click on the green triangle to run the indexer.

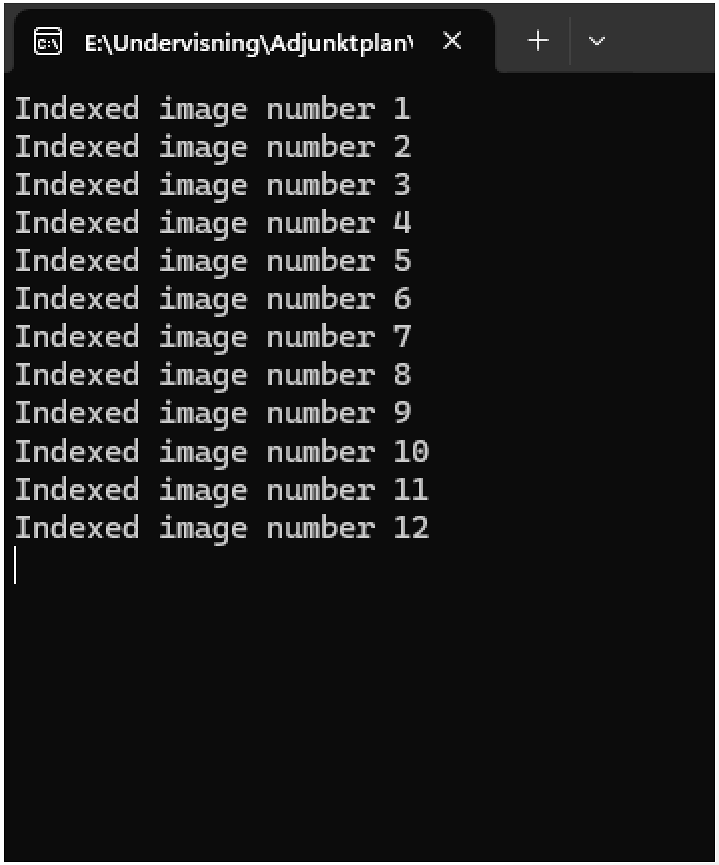


- If it doesen’t work as expected, get my attention and I will take a look at your setup.

Modular RAGs

- If everything is working, you should see the indexer crawling the Azure Blob store image container:

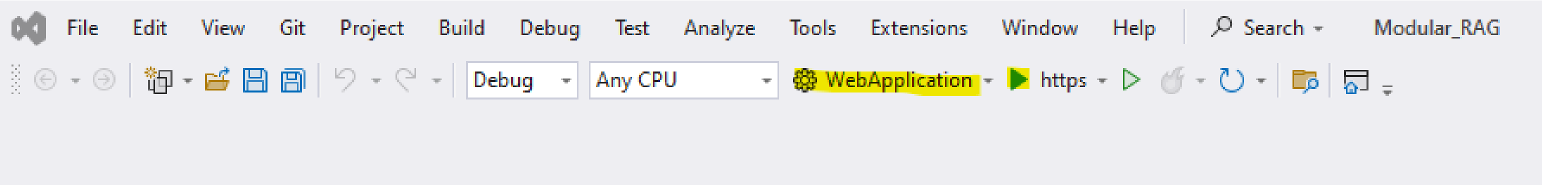
There is a delay of three seconds between each index entry produced, because the free tiers of the services we are using have a limit of 20 requests per minute. Let the indexer do its job in the foreground and leave your machine to it until 60 images have been processed.



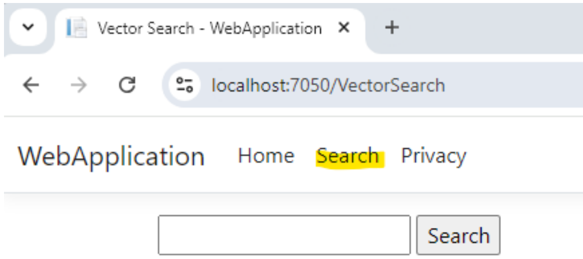
```
Indexed image number 1
Indexed image number 2
Indexed image number 3
Indexed image number 4
Indexed image number 5
Indexed image number 6
Indexed image number 7
Indexed image number 8
Indexed image number 9
Indexed image number 10
Indexed image number 11
Indexed image number 12
```

Modular RAGs

- Once the indexer is done, choose the Web Application from the top menu, and click on the green triangle to start the Web App.

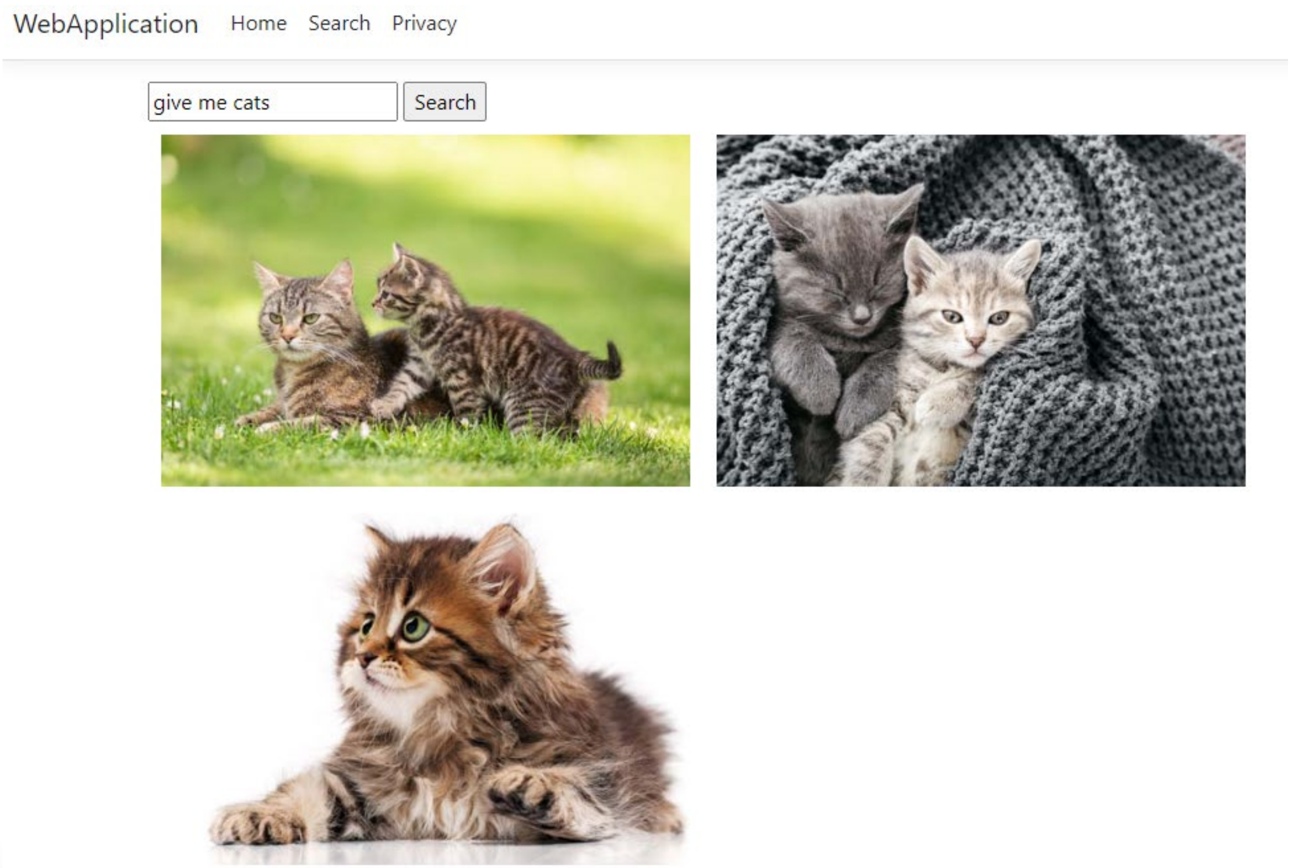


- Click on ‘Search’ to enter a search term for an animal.



Modular RAGs

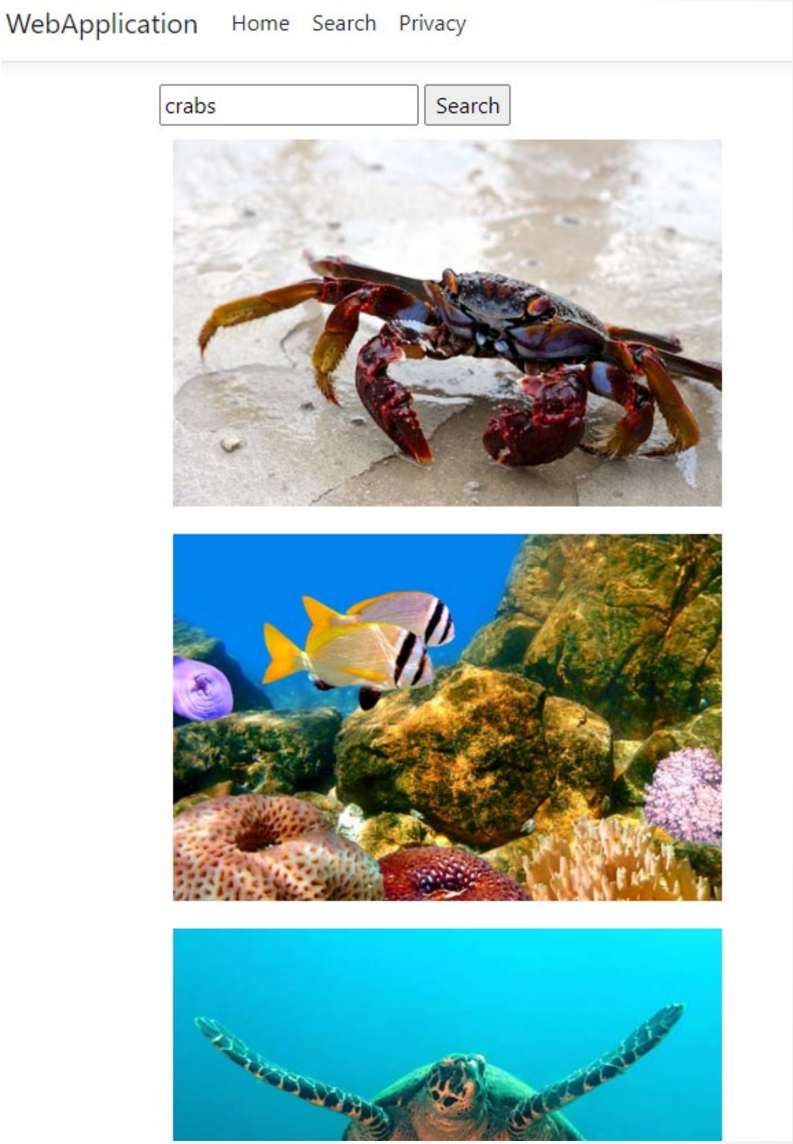
- Here, I have searched for cats:



give me cats

Modular RAGs

- .. and here I have searched for crabs. The vector search will return the highest probability of a match first (the only picture of a crab in the image collection) and it will ‘hallucinate’ the closest match it thinks it can find for the next two images.



crabs

Modular RAG's

- Learn more about Azure AI Services
- Microsoft Learn AI Services landing page: https://learn.microsoft.com/en-us/azure/ai-services/
- AI Search: https://learn.microsoft.com/en-us/azure/search/
- Computer Vision: https://learn.microsoft.com/en-us/azure/ai-services/computer-vision/