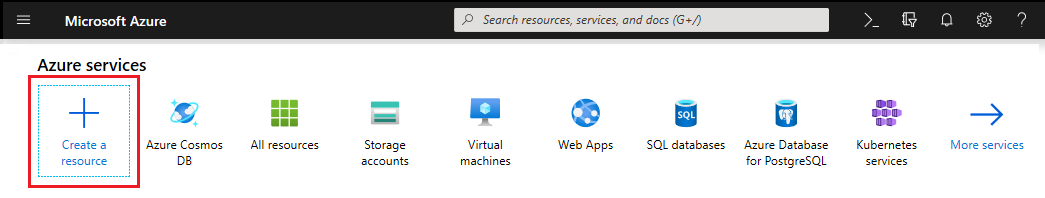
**Provision an Azure AI Search resource**

[**Azure AI Search**](https://learn.microsoft.com/en-us/azure/search/search-what-is-azure-search) is an information retrieval platform for the enterprise. It supports traditional search and conversational AI-driven search for "chat with your data" experiences over your proprietary content.

**Find the Azure AI Search offering**

1. Sign in to the [Azure portal](https://portal.azure.com/).
2. Select (**Create Resource"**) in the top-left corner.
3. Use the search bar to find "Azure AI Search\*.

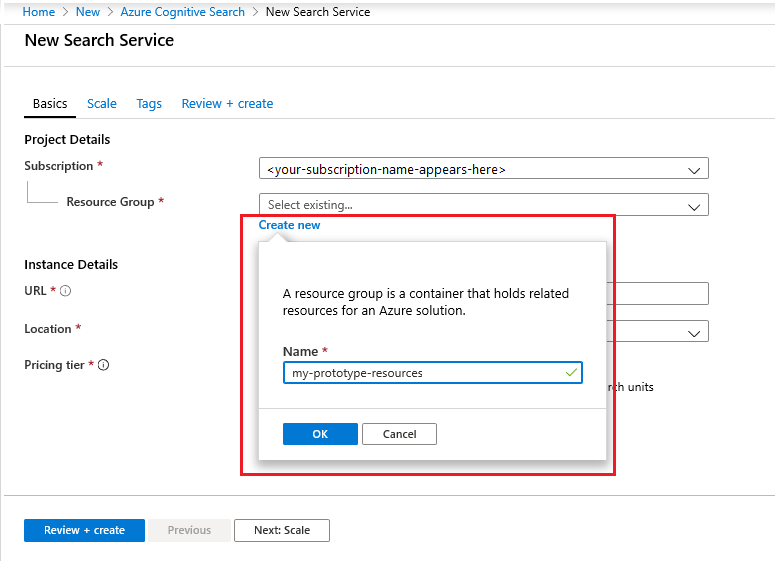
[](https://learn.microsoft.com/en-us/azure/search/media/search-create-service-portal/find-search3.png#lightbox)

**Choose a subscription**

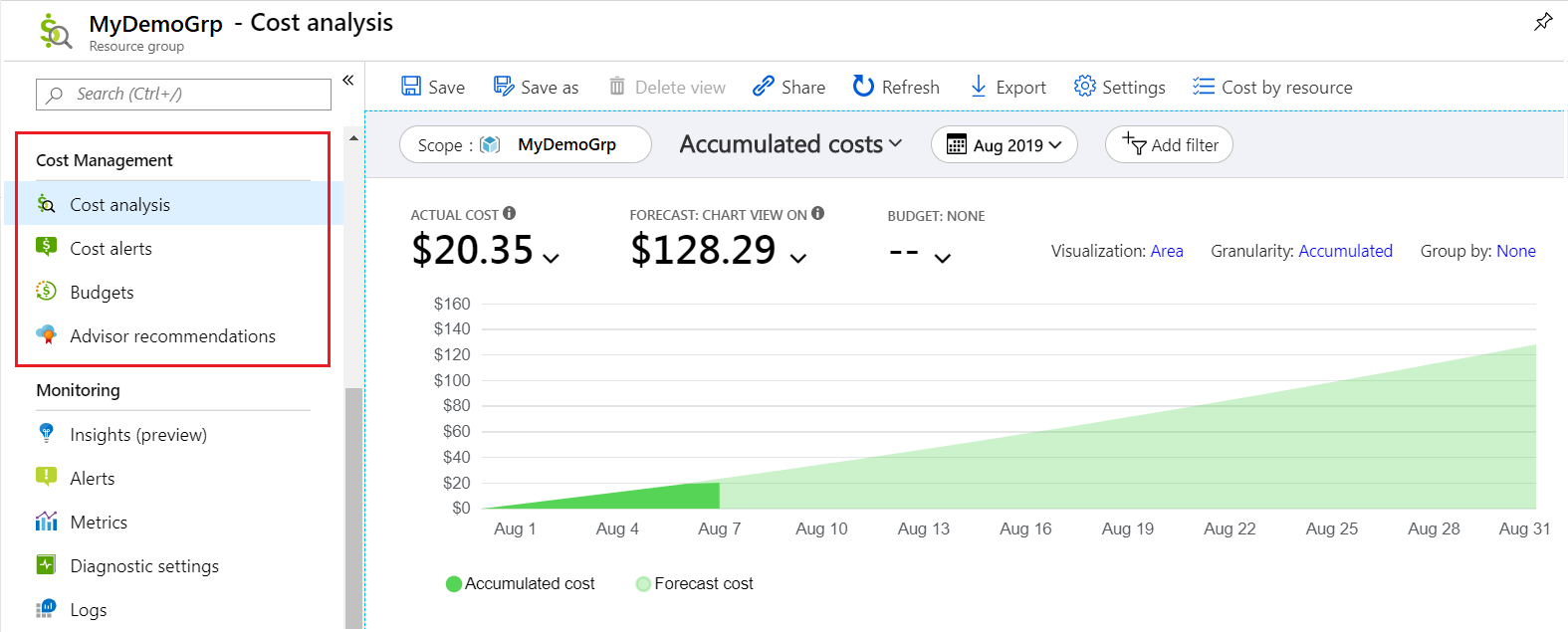
If you have more than one subscription, choose one for your search service. If you're implementing [customer-managed encryption](https://learn.microsoft.com/en-us/azure/search/search-security-manage-encryption-keys) or if you use other features that depend on managed service identities for [external data access](https://learn.microsoft.com/en-us/azure/search/search-indexer-securing-resources), choose the same subscription as the one used for Azure Key Vault or other services for which managed identities are used.

**Set a resource group**

A resource group is a container that holds related resources for your Azure solution. It's useful for consolidating same-solution resources, monitoring costs, and for checking the creation date of your search service.

[](https://learn.microsoft.com/en-us/azure/search/media/search-create-service-portal/new-resource-group.png#lightbox)

Over time, you can track current and projected costs all-up or you can view charges for individual resources. The following screenshot shows the kind of cost information you can expect to see when you combine multiple resources into one group.

[](https://learn.microsoft.com/en-us/azure/search/media/search-create-service-portal/resource-group-cost-management.png#lightbox)

**Tip**

Resource groups simplify cleanup because deleting a resource group deletes everything within it.

**Name the service**

In Instance Details, provide a service name in the **URL** field. The name is part of the endpoint against which API calls are issued: https://your-service-name.search.windows.net. For example, if you want the endpoint to be https://myservice.search.windows.net, you would enter myservice.

Service name requirements:

* Unique within the search.windows.net namespace
* Between 2-60 characters in length
* Consist of lowercase letters, digits, or dashes (-)
* Don't use dashes in the first two characters or as the last single character
* Don't use consecutive dashes anywhere

**Tip :** If you have multiple search services, it helps to include the region (or location) in the service name as a naming convention. A name like **mysearchservice-westus** can save you a trip to the properties page when deciding how to combine or attach resources.

**Choose a region**

**Important :** Due to high demand, Azure AI Search is currently unavailable for new instances in some regions.

If you use multiple Azure services, putting all of them in the same region minimizes or voids bandwidth charges. There are no charges for data egress among same-region services.

Generally, choose a region near you, unless the following considerations apply:

* Your nearest region is [at capacity](https://learn.microsoft.com/en-us/azure/search/search-sku-tier#region-availability-by-tier). One advantage to using the Azure portal for resource setup is that it provides only those regions and tiers that are available.
* You want to use integrated data chunking and vectorization or built-in skills for AI enrichment. Integrated operations have region requirements.
* You want to use Azure Storage for indexer-based indexing or you need to store application data that isn't in an index. Debug session state, enrichment caches, and knowledge stores are Azure AI Search features that have a dependency on Azure Storage. The region you choose for Azure Storage has implications for network security. Specifically, if you're setting up a firewall, you should place the resources in *separate regions*. For more information, see [Outbound connections from Azure AI Search to Azure Storage](https://learn.microsoft.com/en-us/azure/search/search-indexer-securing-resources).

**Checklist for choosing a region**

1. Is Azure AI Search available in a nearby region? Check the [supported regions list](https://learn.microsoft.com/en-us/azure/search/search-region-support).
2. Do you have a specific tier in mind? Check [region availability by tier](https://learn.microsoft.com/en-us/azure/search/search-sku-tier#region-availability-by-tier).
3. Do you have business continuity and disaster recovery (BCDR) requirements? Create two or more search services in [regional pairs](https://learn.microsoft.com/en-us/azure/reliability/cross-region-replication-azure#azure-paired-regions) within [availability zones](https://learn.microsoft.com/en-us/azure/search/search-reliability#availability-zones). For example, if you're operating in North America, you might choose East US and West US, or North Central US and South Central US, for each search service.
4. Do you need [AI enrichment](https://learn.microsoft.com/en-us/azure/search/cognitive-search-concept-intro), [integrated data chunking and vectorization](https://learn.microsoft.com/en-us/azure/search/vector-search-integrated-vectorization), or [multimodal image search](https://learn.microsoft.com/en-us/azure/search/search-get-started-portal-image-search)? Azure AI Search, Azure OpenAI, and Azure AI multiservice must coexist in the same region.
   * Start with [Azure OpenAI regions](https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models#model-summary-table-and-region-availability) because it has the most variability. Azure OpenAI provides embedding models and chat models for RAG and integrated vectorization.
   * Check [Azure AI Search regions](https://learn.microsoft.com/en-us/azure/search/search-region-support) for a match to your Azure OpenAI region. If you're using OCR, entity recognition, or other skills backed by Azure AI, the **AI Integration** column indicates whether Azure AI multiservice is in the same region as Azure AI Search.
   * Check [multimodal embedding regions](https://learn.microsoft.com/en-us/azure/ai-services/computer-vision/overview-image-analysis#region-availability) for multimodal APIs and image search. This API is accessed through an Azure AI multiservice account, but it's available in fewer regions than Azure AI multiservice in general.

**Regions with the most overlap**

Currently, the following regions offer cross-region among all three services (Azure AI Search, Azure OpenAI, Azure AI Vision multimodal). This list isn't definitive, and there might be more choices beyond the regions listed here depending on the tier. Also, region status can change quickly, so be sure to confirm region choice before installing.

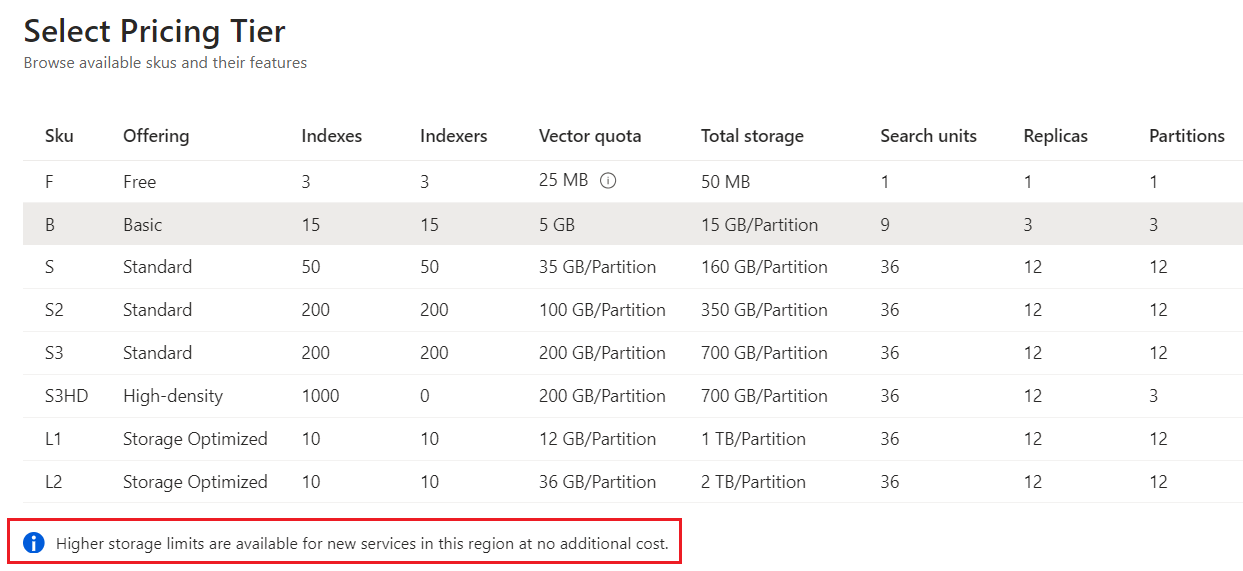
* **Americas**: West US
* **Europe**: France Central, North Europe, Sweden Central

**Choose a tier**

Azure AI Search is offered in [multiple pricing tiers](https://azure.microsoft.com/pricing/details/search/): Free, Basic, Standard, or Storage Optimized. Each tier has its own [capacity and limits](https://learn.microsoft.com/en-us/azure/search/search-limits-quotas-capacity). There are also several features that are tier-dependent.

Review the [tier descriptions](https://learn.microsoft.com/en-us/azure/search/search-sku-tier) for computing characteristics, [feature availability](https://learn.microsoft.com/en-us/azure/search/search-sku-tier#feature-availability-by-tier), and [region availability](https://learn.microsoft.com/en-us/azure/search/search-sku-tier#region-availability-by-tier).

Basic and Standard are the most common choices for production workloads, but many customers start with the Free service. Among the billable tiers, key differences are partition size and speed, and limits on the number of objects you can create.

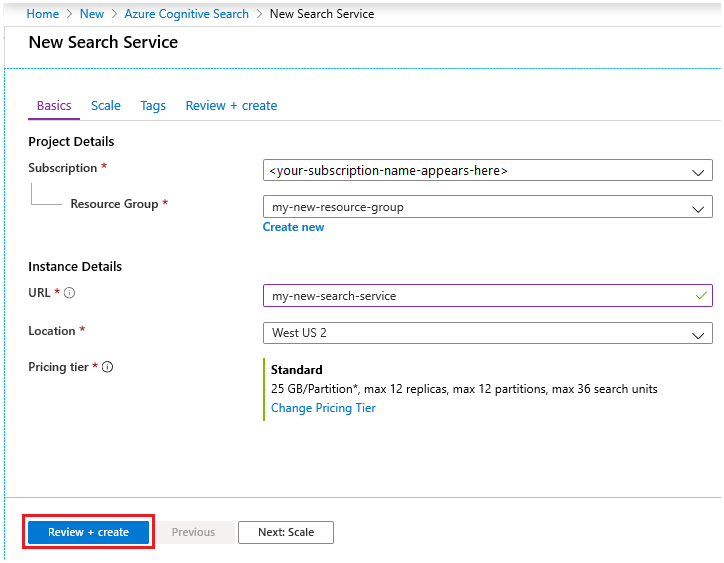
[](https://learn.microsoft.com/en-us/azure/search/media/search-create-service-portal/select-pricing-tier.png#lightbox)

Search services created after April 3, 2024 have larger partitions and higher vector quotas at every billable tier.

Remember, a pricing tier can't be changed once the service is created. If you need a higher or lower tier, you should re-create the service.

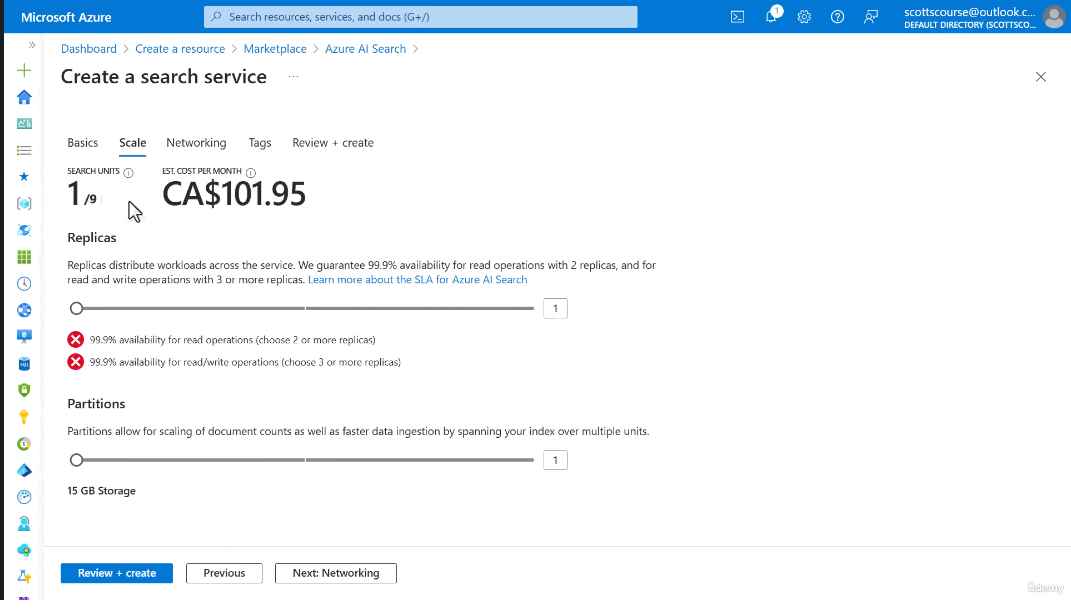
**Create your service**

After you've provided the necessary inputs, go ahead and create the service.

[](https://learn.microsoft.com/en-us/azure/search/media/search-create-service-portal/new-service3.png#lightbox)

A screenshot of a computer

Description automatically generated



A screenshot of a computer

Description automatically generated

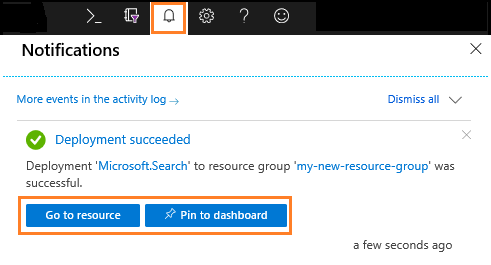
A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Your service is deployed within minutes. You can monitor progress through Azure notifications. Consider pinning the service to your dashboard for easy access in the future.

[](https://learn.microsoft.com/en-us/azure/search/media/search-create-service-portal/monitor-notifications.png#lightbox)

A screenshot of a computer

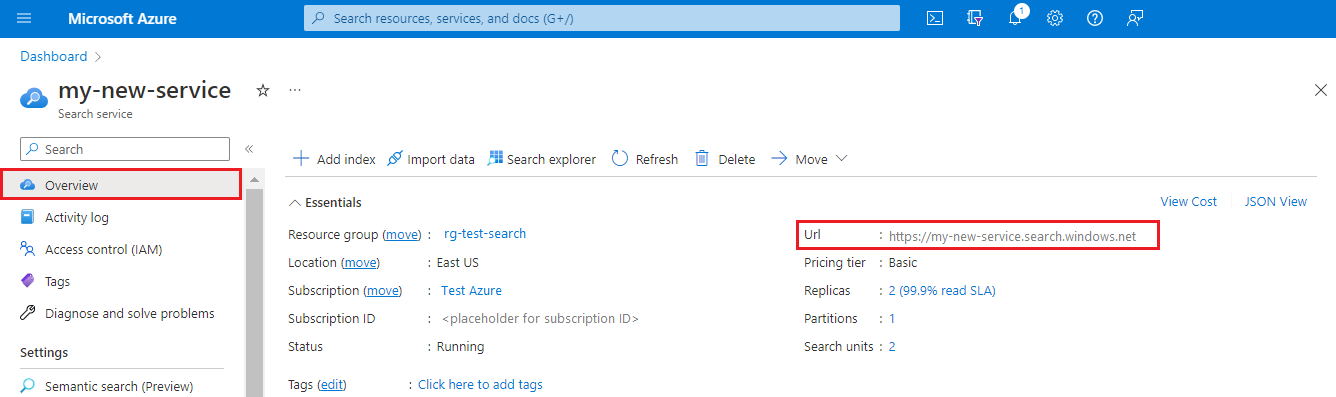
Description automatically generated

**Configure authentication**

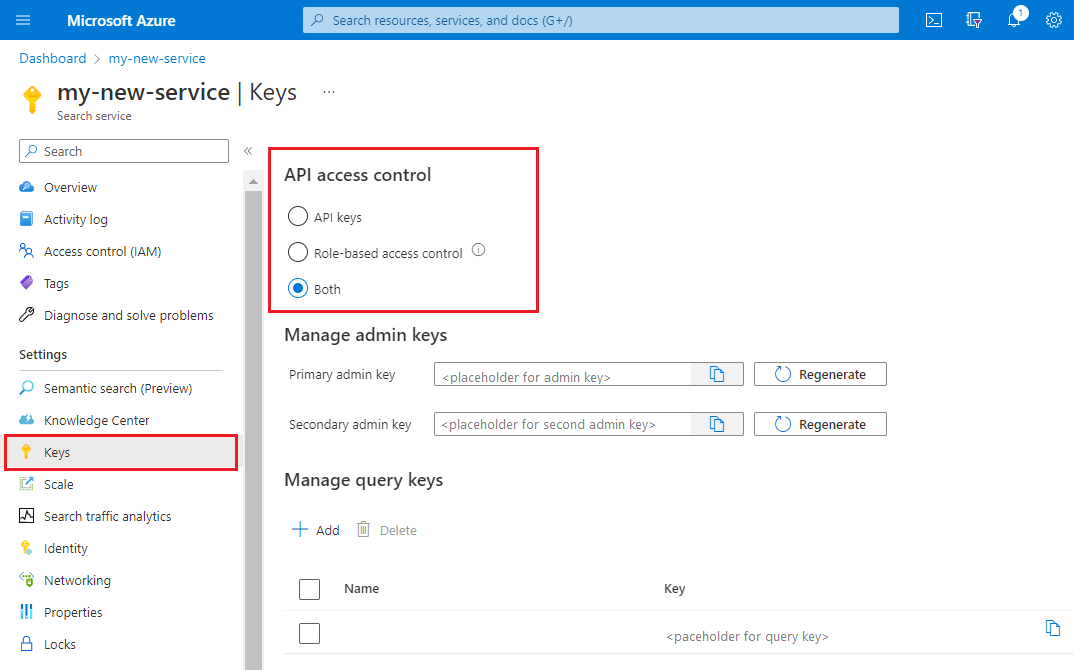
Unless you're using the Azure portal, programmatic access to your new service requires that you provide the URL endpoint and an authenticated connection. You can use either or both of these options:

* [Connect using key-based authentication](https://learn.microsoft.com/en-us/azure/search/search-security-api-keys)
* [Connect using Azure roles](https://learn.microsoft.com/en-us/azure/search/search-security-rbac)

1. When setting up a programmatic connection, you need the search service endpoint. On the **Overview** page, locate and copy the URL endpoint on the right side of the page.

[](https://learn.microsoft.com/en-us/azure/search/media/search-create-service-portal/get-endpoint.png#lightbox)

1. To set authentication options, use the **Keys** page. Most quickstarts and tutorials use API keys for simplicity, but if you're setting up a service for production workloads, consider using Azure roles. You can copy keys from this page.

[](https://learn.microsoft.com/en-us/azure/search/media/search-create-service-portal/set-authentication-options.png#lightbox)

An endpoint and key aren't needed for portal-based tasks. the Azure portal is already linked to your Azure AI Search resource with admin rights. For a portal walkthrough, start with [Quickstart: Create an Azure AI Search index in the Azure portal](https://learn.microsoft.com/en-us/azure/search/search-get-started-portal).

**Scale your service**

After a search service is provisioned, you can [scale it to meet your needs](https://learn.microsoft.com/en-us/azure/search/search-limits-quotas-capacity). On a billable tier, you can scale the service in two dimensions: replicas and partitions. For the free service, scale up isn't available and replica and partition configuration isn't offered.

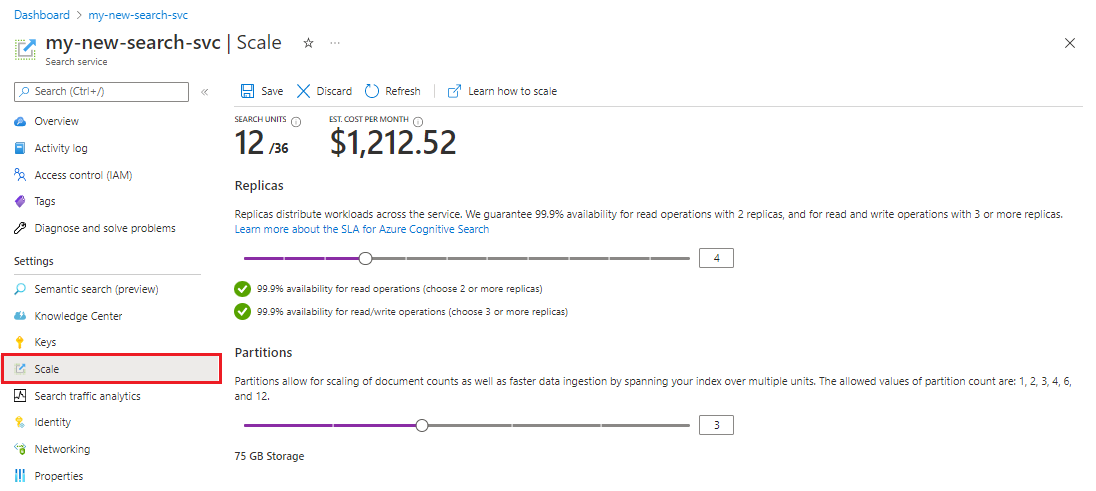
***Partitions*** allow your service to store and search through more documents.

***Replicas*** allow your service to handle a higher load of search queries.

Adding resources increases your monthly bill. The [pricing calculator](https://azure.microsoft.com/pricing/calculator/) can help you understand the billing ramifications of adding resources. Remember that you can adjust resources based on load. For example, you might increase resources to create a full initial index, and then reduce resources later to a level more appropriate for incremental indexing.

**Important :** A service must have [**2 replicas for read-only SLA and 3 replicas for read/write SLA**](https://azure.microsoft.com/support/legal/sla/search/v1_0/).

1. Go to your search service page in the Azure portal.
2. In the left-navigation pane, select **Settings** > **Scale**.
3. Use the slidebar to add resources of either type.

[](https://learn.microsoft.com/en-us/azure/search/media/search-create-service-portal/settings-scale.png#lightbox)

**When to add a second service**

Most customers use just one service provisioned at a tier [sufficient for expected load](https://learn.microsoft.com/en-us/azure/search/search-capacity-planning). One service can host multiple indexes, subject to the [maximum limits of the tier you select](https://learn.microsoft.com/en-us/azure/search/search-limits-quotas-capacity#index-limits), with each index isolated from another. In Azure AI Search, requests can only be directed to one index, minimizing the chance of accidental or intentional data retrieval from other indexes in the same service.

Although most customers use just one service, service redundancy might be necessary if operational requirements include the following:

* [Business continuity and disaster recovery (BCDR)](https://learn.microsoft.com/en-us/azure/reliability/cross-region-replication-azure). Azure AI Search doesn't provide instant failover if there's an outage.
* [Multitenant architectures](https://learn.microsoft.com/en-us/azure/search/search-modeling-multitenant-saas-applications) sometimes call for two or more services.
* Globally deployed applications might require search services in each geography to minimize latency.

**Note :** In Azure AI Search, you cannot segregate indexing and querying operations; thus, you would never create multiple services for segregated workloads. An index is always queried on the service in which it was created (you cannot create an index in one service and copy it to another).

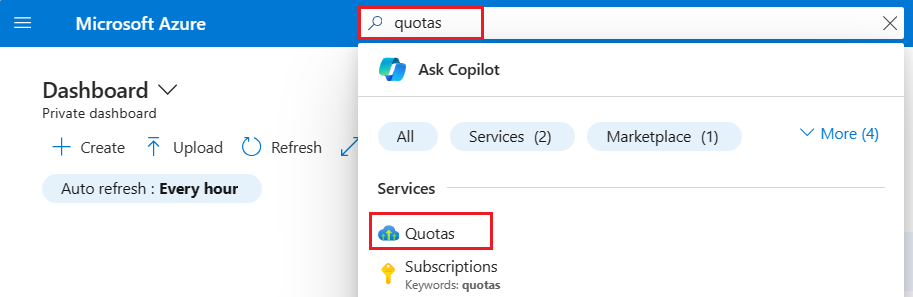
A second service isn't required for high availability. High availability for queries is achieved when you use two or more replicas in the same service. Replica updates are sequential, which means at least one is operational when a service update is rolled out. For more information about uptime, see [Service Level Agreements](https://azure.microsoft.com/support/legal/sla/search/v1_0/).

**Add more services to a subscription**

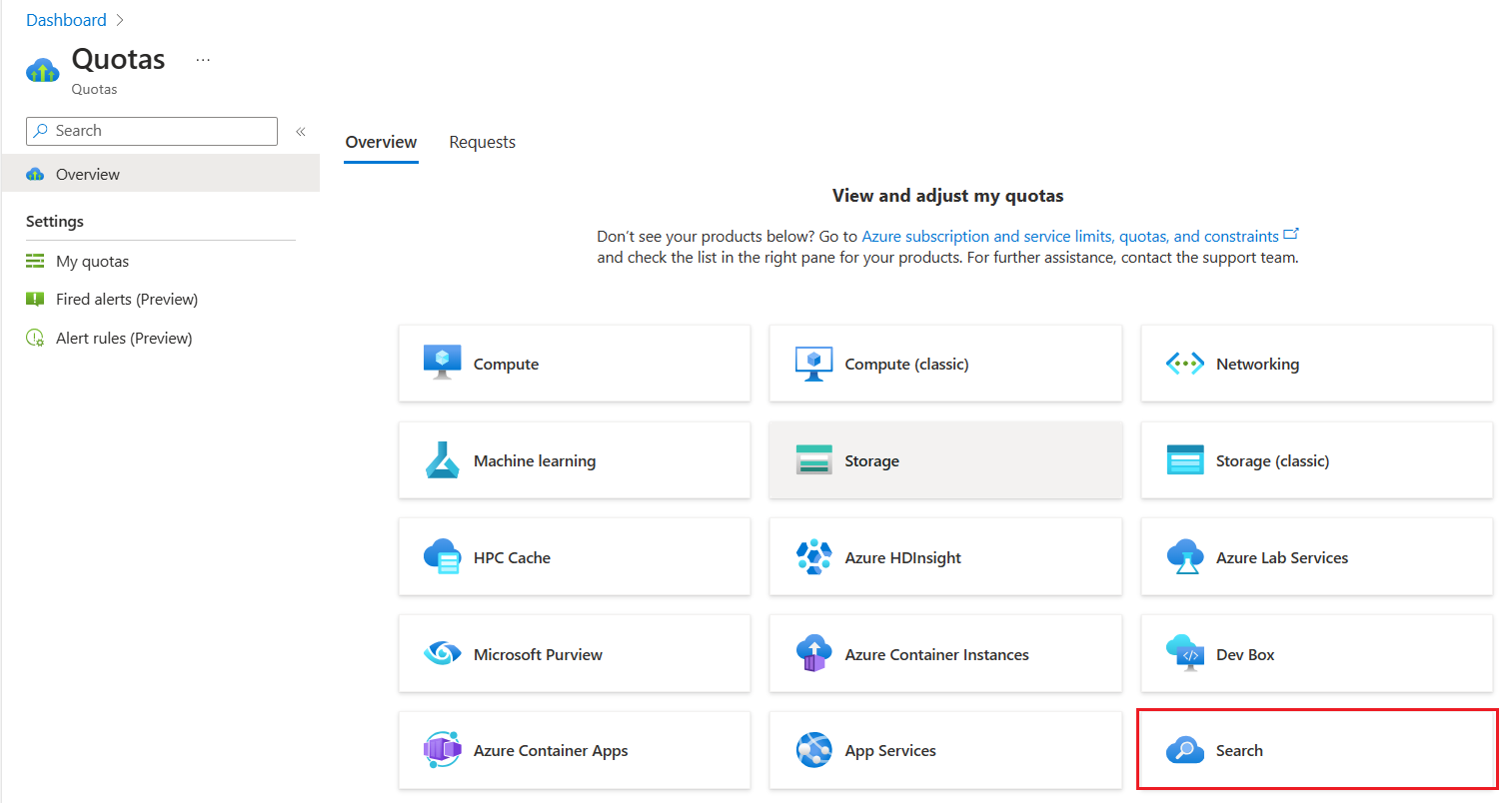
Azure AI Search restricts the [number of search services](https://learn.microsoft.com/en-us/azure/search/search-limits-quotas-capacity#subscription-limits) you can initially create in a subscription. If you exhaust your maximum limit, you can request more quota.

You must have Owner or Contributor permissions on the subscription to request quota. Depending on region and datacenter capacity, you can automatically request more quota to add services to your subscription. If the request fails, you should either decrease the number or file a support ticket. For a large increase in quota, such as more than 30 extra services, you should expect a one-month turnaround.

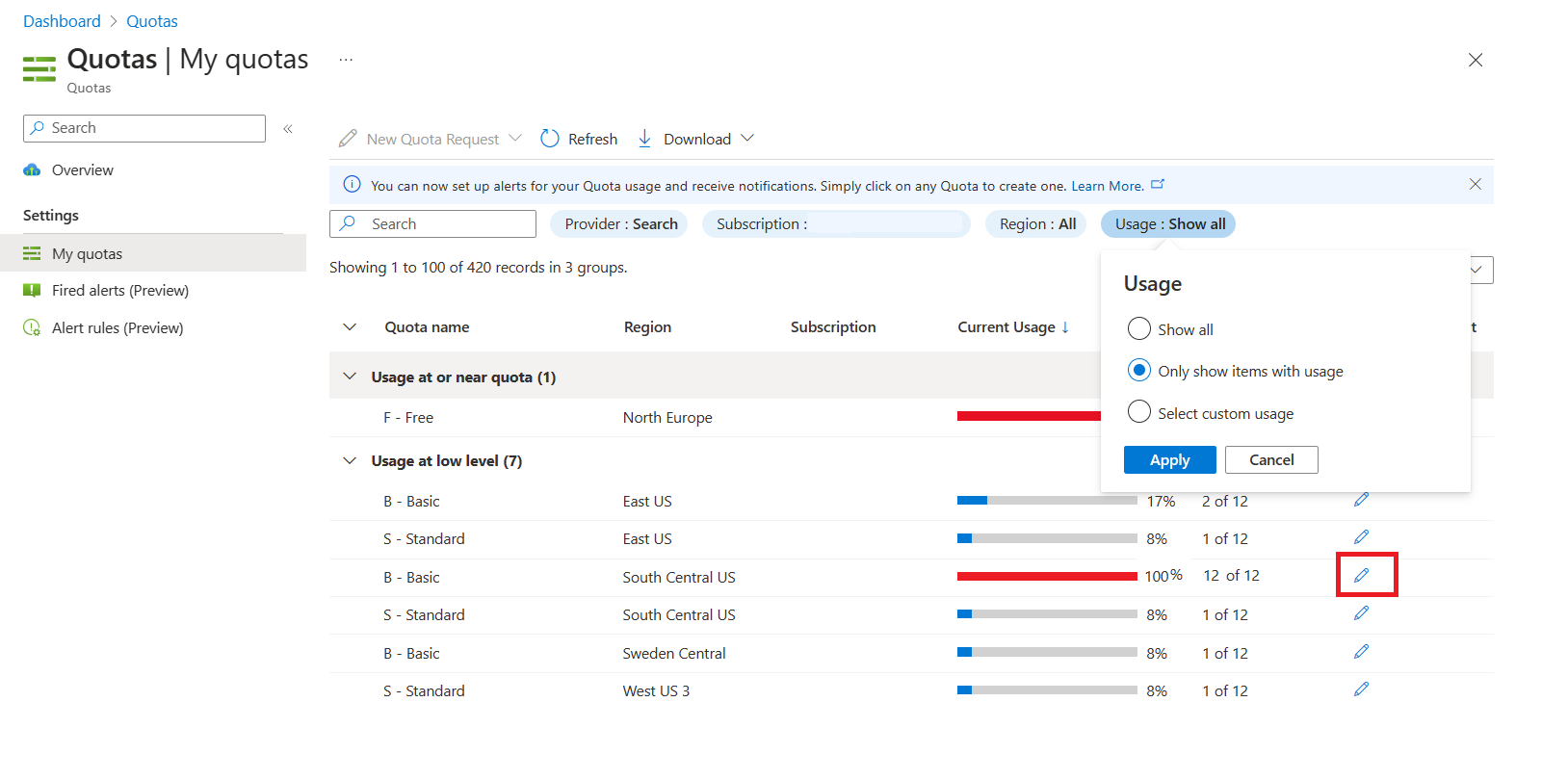
1. Sign in to the Azure portal, search for "quotas" in your dashboard, and then select the **Quotas** service.

[](https://learn.microsoft.com/en-us/azure/search/media/search-create-service-portal/quota-search.png#lightbox)

1. In the Quota's Overview page, select **Search**.

[](https://learn.microsoft.com/en-us/azure/search/media/search-create-service-portal/quota-overview-page.png#lightbox)

1. Set filters so that you can review existing quota for search services in the current subscription. We recommend filtering by usage.
2. Find the region and tier that needs more quota and select the **Edit** pencil icon to begin your request.

[](https://learn.microsoft.com/en-us/azure/search/media/search-create-service-portal/quota-pencil-edit.png#lightbox)

1. In **Quota details**, specify the location, tier, and a new limit for your subscription quota. None of the values can be empty. The new limit must be greater than the current limit. If regional capacity is constrained, your request won't be automatically approved. In this scenario, an incident report is generated on your behalf for investigation and resolution.
2. Submit the request.
3. Monitor notifications in the Azure portal for status updates on the new limit. Most requests are approved within 24 hours.