Service limits in Azure Cognitive Search

10/14/2020 • 11 minutes to read • +6

In this article

Maximum limits on storage, workloads, and quantities of indexes and other objects depend on whether you provision Azure Cognitive Search at **Free**, **Basic**, **Standard**, or **Storage Optimized** pricing tiers.

- **Free** is a multi-tenant shared service that comes with your Azure subscription.
- **Basic** provides dedicated computing resources for production workloads at a smaller scale, but shares some networking infrastructure with other tenants.
- **Standard** runs on dedicated machines with more storage and processing capacity at every level. Standard comes in four levels: S1, S2, S3, and S3 HD. S3 High Density (S3 HD) is engineered for multi-tenancy and large quantities of small

indexes (three thousand indexes per service). S3 HD does not provide the indexer feature and data ingestion must leverage APIs that push data from source to index.

- **Storage Optimized** runs on dedicated machines with more total storage, storage bandwidth, and memory than **Standard**. This tier targets large, slow-changing indexes. Storage Optimized comes in two levels: L1 and L2.

Subscription limits

You can create multiple services within a subscription. Each one can be provisioned at a specific tier. You're limited only by the number of services allowed at each tier. For example, you could create up to 12 services at the Basic tier and another 12 services at the S1 tier within the same subscription. For more information about tiers, see Choose an SKU or tier for Azure Cognitive Search.

Maximum service limits can be raised upon request. If you need more services within the same subscription, contact Azure Support.

Resource	Free[1]	Basic	S1	S2	S3	S3 HD	L1	L2
Maximum services	1	16	16	8	6	6	6	6
Maximum scale in search units (SU)[2]	N/A	3 SU	36 SU	36 SU	36 SU	36 SU	36 SU	36 SU

[1] Free is based on shared, not dedicated, resources. Scale-up is not supported on shared resources.

[2] Search units are billing units, allocated as either a *replica* or a *partition*. You need both resources for storage, indexing, and query operations. To learn more about SU computations, see Scale resource levels for query and index workloads.

Storage limits

A search service is constrained by disk space or by a hard limit on the maximum number of indexes or indexers, whichever comes first. The following table documents storage limits. For maximum object limits, see Limits by resource.

Resource	Free	Basic[1]	S1	S2	S3	S3 HD	L1	L2
Service level agreement (SLA)[2]	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Storage per partition	50 MB	2 GB	25 GB	100 GB	200 GB	200 GB	1 TB	2 TB
Partitions per service	N/A	1	12	12	12	3	12	12
Partition size	N/A	2 GB	25 GB	100 GB	200 GB	200 GB	1 TB	2 TB
Replicas	N/A	3	12	12	12	12	12	12

[1] Basic has one fixed partition. Additional search units can be used to add replicas for larger query volumes.

[2] Service level agreements are in effect for billable services on dedicated resources. Free services and preview features have no SLA. For billable services, SLAs take effect when you provision sufficient redundancy for your service. Two or more replicas are required for query (read) SLAs. Three or more replicas are required for query and indexing (read-write) SLAs. The number of partitions isn't an SLA consideration.

Index limits

Resource	Free	Basic [1]	S1	S2	S3	S3 HD	L1	L2
Maximum indexes	3	5 or 15	50	200	200	1000 per partition or 3000 per service	10	10

Resource	Free	Basic [1]	S1	S2	S3	S3 HD	L1	L2
Maximum simple fields per index	1000	100	1000	1000	1000	1000	1000	1000
Maximum complex collection fields per index	40	40	40	40	40	40	40	40
Maximum elements across all complex collections per document [2]	3000	3000	3000	3000	3000	3000	3000	3000
Maximum depth of complex fields	10	10	10	10	10	10	10	10
Maximum suggesters per index	1	1	1	1	1	1	1	1
Maximum scoring profiles per index	100	100	100	100	100	100	100	100
Maximum functions per profile	8	8	8	8	8	8	8	8

[1] Basic services created before December 2017 have lower limits (5 instead of 15) on indexes. Basic tier is the only SKU with a lower limit of 100 fields per index.

[2] An upper limit exists for elements because having a large number of them significantly increases the storage required for your index. An element of a complex collection is defined as a member of that collection. For example, assume a Hotel document with a Rooms complex collection, each room in the Rooms collection is considered an element. During indexing, the indexing engine can safely process a maximum of 3000 elements across the document as a whole. This limit was introduced in `api-version=2019-05-06` and applies to complex collections only, and not to string collections or to complex fields.

Document limits

As of October 2018, there are no longer any document count limits for any new service created at any billable tier (Basic, S1, S2, S3, S3 HD) in any region. Older services created prior to October 2018 may still be subject to document count limits.

To determine whether your service has document limits, use the GET Service Statistics REST API. Document limits are reflected in the response, with `null` indicating no limits.

> ⓘ **Note**
>
> Although there are no document limits imposed by the service, there is a shard limit of approximately 24 billion documents per index on Basic, S1, S2, and S3 search services. For S3 HD, the shard limit is 2 billion documents per index. Each element of a complex collection counts as a separate document in terms of shard limits.

Document size limits per API call

The maximum document size when calling an Index API is approximately 16 megabytes.

Document size is actually a limit on the size of the Index API request body. Since you can pass a batch of multiple documents to the Index API at once, the size limit realistically depends on how many documents are in the batch. For a batch with a single document, the maximum document size is 16 MB of JSON.

When estimating document size, remember to consider only those fields that can be consumed by a search service. Any binary or image data in source documents should be omitted from your calculations.

Indexer limits

Maximum running times exist to provide balance and stability to the service as a whole, but larger data sets might need more indexing time than the maximum allows. If an indexing job cannot complete within the maximum time allowed, try running it on a schedule. The scheduler keeps track of indexing status. If a scheduled indexing job is interrupted for any reason, the indexer can pick up where it last left off at the next scheduled run.

Resource	Free [1]	Basic [2]	S1	S2	S3	S3 HD [3]	L1	L2
Maximum indexers	3	5 or 15	50	200	200	N/A	10	10
Maximum datasources	3	5 or 15	50	200	200	N/A	10	10
Maximum skillsets [4]	3	5 or 15	50	200	200	N/A	10	10
Maximum indexing load per invocation	10,000 documents	Limited only by maximum documents	Limited only by maximum documents	Limited only by maximum documents	Limited only by maximum documents	N/A	No limit	No limit
Minimum schedule	5 minutes	5 minutes	5 minutes	5 minutes	5 minutes	5 minutes	5 minutes	5 minutes
Maximum running time	1-3 minutes	24 hours	24 hours	24 hours	24 hours	N/A	24 hours	24 hours
Maximum running time for indexers with a skillset [5]	3-10 minutes	2 hours	2 hours	2 hours	2 hours	N/A	2 hours	2 hours
Blob indexer: maximum blob size, MB	16	16	128	256	256	N/A	256	256

Resource	Free [1]	Basic [2]	S1	S2	S3	S3 HD [3]	L1	L2
Blob indexer: maximum characters of content extracted from a blob	32,000	64,000	4 million	8 million	16 million	N/A	4 million	4 million

[1] Free services have indexer maximum execution time of 3 minutes for blob sources and 1 minute for all other data sources. For AI indexing that calls into Cognitive Services, free services are limited to 20 free transactions per day, where a transaction is defined as a document that successfully passes through the enrichment pipeline.

[2] Basic services created before December 2017 have lower limits (5 instead of 15) on indexers, data sources, and skillsets.

[3] S3 HD services do not include indexer support.

[4] Maximum of 30 skills per skillset.

[5] AI enrichment and image analysis are computationally intensive and consume disproportionate amounts of available processing power. Running time for these workloads has been shortened to give other jobs in the queue more opportunity to run.

ⓘ **Note**

As stated in the Index limits, indexers will also enforce the upper limit of 3000 elements across all complex collections per document starting with the latest GA API version that supports complex types (`2019-05-06`) onwards. This means that if you've created your indexer with a prior API version, you will not be subject to this limit. To preserve maximum compatibility, an indexer that was created with a prior API version and then updated with an API version `2019-05-06` or later, will still be **excluded** from the limits. Customers should be aware of the adverse impact of having very large

> complex collections (as stated previously) and we highly recommend creating any new indexers with the latest GA API version.

Shared private link resource limits

Indexers can access other Azure resources over private endpoints managed via the shared private link resource API. This section describes the limits associated with this capability.

Resource	Free	Basic	S1	S2	S3	S3 HD	L1	L2
Private endpoint indexer support	No	Yes	Yes	Yes	Yes	No	Yes	Yes
Private endpoint support for indexers with a skillset[1]	No	No	No	Yes	Yes	No	Yes	Yes
Maximum private endpoints	N/A	10 or 30	100	400	400	N/A	20	20
Maximum distinct resource types[2]	N/A	4	7	15	15	N/A	4	4

[1] AI enrichment and image analysis are computationally intensive and consume disproportionate amounts of available processing power. For this reason, private connections are disabled on lower tiers to avoid an adverse impact on the performance and stability of the search service itself.

[2] The number of distinct resource types are computed as the number of unique `groupId` values used across all shared private link resources for a given search service, irrespective of the status of the resource.

Synonym limits

Maximum number of synonym maps varies by tier. Each rule can have up to 20 expansions, where an expansion is an equivalent term. For example, given "cat", association with "kitty", "feline", and "felis" (the genus for cats) would count as 3 expansions.

Resource	Free	Basic	S1	S2	S3	S3-HD	L1	L2
Maximum synonym maps	3	3	5	10	20	20	10	10
Maximum number of rules per map	5000	20000	20000	20000	20000	20000	20000	20000

Queries per second (QPS)

QPS estimates must be developed independently by every customer. Index size and complexity, query size and complexity, and the amount of traffic are primary determinants of QPS. There is no way to offer meaningful estimates when such factors are unknown.

Estimates are more predictable when calculated on services running on dedicated resources (Basic and Standard tiers). You can estimate QPS more closely because you have control over more of the parameters. For guidance on how to approach estimation, see Azure Cognitive Search performance and optimization.

For the Storage Optimized tiers (L1 and L2), you should expect a lower query throughput and higher latency than the Standard tiers.

Data limits (AI enrichment)

An AI enrichment pipeline that makes calls to a Text Analytics resource for entity recognition, key phrase extraction, sentiment analysis, language detection, and personal-information detection is subject to data limits. The maximum size of a record should be 50,000 characters as measured by String.Length. If you need to break up your data before sending it to the sentiment analyzer, use the Text Split skill.

Throttling limits

Search query and indexing requests are throttled as the system approaches peak capacity. Throttling behaves differently for different APIs. Query APIs (Search/Suggest/Autocomplete) and indexing APIs throttle dynamically based on the load on the service. Index APIs have static request rate limits.

Static rate request limits for operations related to an index:

- List Indexes (GET /indexes): 5 per second per search unit
- Get Index (GET /indexes/myindex): 10 per second per search unit
- Create Index (POST /indexes): 12 per minute per search unit
- Create or Update Index (PUT /indexes/myindex): 6 per second per search unit
- Delete Index (DELETE /indexes/myindex): 12 per minute per search unit

API request limits

- Maximum of 16 MB per request [1]
- Maximum 8 KB URL length
- Maximum 1000 documents per batch of index uploads, merges, or deletes
- Maximum 32 fields in $orderby clause
- Maximum search term size is 32,766 bytes (32 KB minus 2 bytes) of UTF-8 encoded text

[1] In Azure Cognitive Search, the body of a request is subject to an upper limit of 16 MB, imposing a practical limit on the contents of individual fields or collections that are not otherwise constrained by theoretical limits (see Supported data types for more information about field composition and restrictions).

API response limits

- Maximum 1000 documents returned per page of search results
- Maximum 100 suggestions returned per Suggest API request

API key limits

API keys are used for service authentication. There are two types. Admin keys are specified in the request header and grant full read-write access to the service. Query keys are read-only, specified on the URL, and typically distributed to client applications.

- Maximum of 2 admin keys per service
- Maximum of 50 query keys per service

Is this page helpful?

Yes No