
Getting Started with Amazon CloudSearch



Getting Started with Amazon CloudSearch

Copyright © 2013 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

The following are trademarks of Amazon Web Services, Inc.: Amazon, Amazon Web Services Design, AWS, Amazon CloudFront, Cloudfront, Amazon DevPay, DynamoDB, ElastiCache, Amazon EC2, Amazon Elastic Compute Cloud, Amazon Glacier, Kindle, Kindle Fire, AWS Marketplace Design, Mechanical Turk, Amazon Redshift, Amazon Route 53, Amazon S3, Amazon VPC. In addition, Amazon.com graphics, logos, page headers, button icons, scripts, and service names are trademarks, or trade dress of Amazon in the U.S. and/or other countries. Amazon's trademarks and trade dress may not be used in connection with any product or service that is not Amazon's, in any manner that is likely to cause confusion among customers, or in any manner that disparages or discredits Amazon.

All other trademarks not owned by Amazon are the property of their respective owners, who may or may not be affiliated with, connected to, or sponsored by Amazon.

Getting Started	1
Step 1: Before You Begin	2
Step 2: Create a Search Domain	2
Step 3: Send Data for Indexing	6
Step 4: Search Your Amazon CloudSearch Domain	9
Step 5: Delete Your Amazon CloudSearch Movies Domain	14

Getting Started with Amazon CloudSearch

Topics

- [Step 1: Before You Begin with Amazon CloudSearch \(p. 2\)](#)
- [Step 2: Create an Amazon CloudSearch Domain \(p. 2\)](#)
- [Step 3: Send Data to Amazon CloudSearch for Indexing \(p. 6\)](#)
- [Step 4: Search Your Amazon CloudSearch Domain \(p. 9\)](#)
- [Step 5: Delete Your Amazon CloudSearch Movies Domain \(p. 14\)](#)

To start searching your data with Amazon CloudSearch, you simply:

- Create and configure a search domain
- Upload and index the data you want to search
- Send search requests to your domain

This tutorial shows you how to get up and running using the AWS Management Console for Amazon CloudSearch. To make it even easier to get started, we've generated a sample data set of over 5,000 popular movie titles that you can download and examine, upload to your own search domain, and submit search queries against to see how Amazon CloudSearch works.

Using the AWS Management Console and the sample movie data, you'll quickly have your own searchable movie database running in Amazon CloudSearch.

To begin, [Get Signed Up](#).

The following video steps through this tutorial and shows how to create your first search domain through the console: [Getting Started with Amazon CloudSearch](#).

Step 1: Before You Begin with Amazon CloudSearch

To use Amazon CloudSearch, you need an Amazon Web Services (AWS) account. Your AWS account enables you to access Amazon CloudSearch and other AWS services, such as Amazon Simple Storage Service (Amazon S3) and Amazon Elastic Compute Cloud (Amazon EC2). As with other AWS services, you pay only for the Amazon CloudSearch resources you use. There are no sign up fees and charges are not incurred until you create a search domain.

If you already have an AWS account, you are automatically signed up for Amazon CloudSearch.

Note

For console access, use your IAM user name and password to sign in to the [AWS Management Console](#) using the [IAM sign-in page](#). IAM lets you securely control access to AWS services and resources in your AWS account. For more information about getting credentials, see [How Do I Get Security Credentials?](#) in the *AWS General Reference*.

To create an AWS account

1. Go to <https://aws.amazon.com> and click **Sign Up Now**.
2. Follow the instructions to sign up. You will need to enter payment information before you can begin using Amazon CloudSearch.

Step 2: Create an Amazon CloudSearch Domain

An Amazon CloudSearch domain encapsulates a collection of data you want to search, the search instances that process your search requests, and a configuration that controls how your data is indexed and searched. You create a separate search domain for each collection of data you want to make searchable. For each domain, you configure indexing options that describe the fields you want to include in your index and how you want to use them, text options that define domain-specific stopwords, stems, and synonyms, rank expressions that you can use to customize how search results are ranked, and access policies that control access to the domains document and search endpoints.

You interact with a search domain to:

- Configure index and search options
- Submit data for indexing
- Perform searches

Each domain has a unique endpoint through which you submit search requests to the domain. For example, the endpoint for a domain called *movies* created in the US East (Northern Virginia) Region might be:

```
search.123456789012-movies.us-east-1.cloudsearch.amazonaws.com
```

When creating a search domain, you specify a unique name for the domain. Domain names must start with a letter or number and be at least 3 and no more than 28 characters long. The allowed characters are: a-z, 0-9, and hyphen (-). By default, new domains are created in the US East (Northern Virginia) Region. To create a domain in another region, you must explicitly specify the region when creating the domain.

To configure the new domain, you need to specify:

Getting Started with Amazon CloudSearch

Step 2: Create a Search Domain

- The index fields you want to be able to search, use as facets, and return in search results.
- Access policies for the domain's document service and search service endpoints.

This tutorial shows you how to create and interact with a domain using the Amazon CloudSearch console. For information about how to use the command line tools and APIs, see [Creating an Amazon CloudSearch Domain](#).

Important

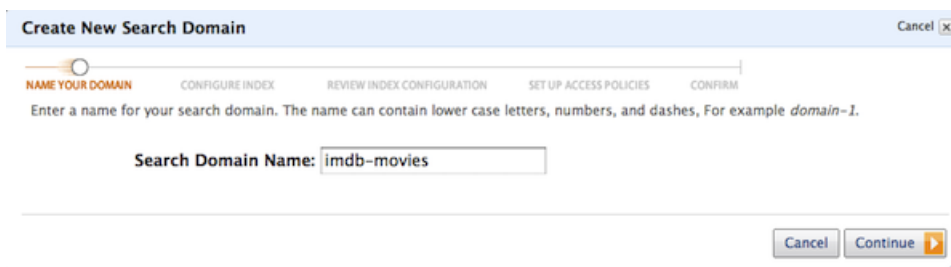
The domain you're about to create will be live and you will incur the standard Amazon CloudSearch usage fees for the domain until you delete it. For more information about Amazon CloudSearch usage rates, go to the [Amazon CloudSearch detail page](#).

To create your movies domain

1. Go to the Amazon CloudSearch console at [Amazon CloudSearch console](#).
2. On the Welcome to Amazon CloudSearch page, click **Create Your First Search Domain**.



3. On the **NAME YOUR DOMAIN** step, enter a name for your new domain and click **Continue**. Domain names must start with a letter or number and be at least 3 and no more than 28 characters. Domain names can contain the following characters: a-z (lower case), 0-9, and - (hyphen). Upper case letters and underscores are not allowed.



4. On the **CONFIGURE INDEX** step, click **Use a predefined configuration**, select **IMDB movies (demo)**, and click **Continue**. You can also automatically configure a search domain by choosing the predefined configuration for the type of data you want to index, or by uploading a sample of your data.

Getting Started with Amazon CloudSearch Step 2: Create a Search Domain

○ Use a predefined configuration

Load configuration for indexing: **IMDB movies (demo)**

- On the **REVIEW INDEX CONFIGURATION** step, review the index fields that will be configured. Five fields are configured automatically for the imdb-movie data: actor, director, genre, title, and year.
 - The actor, director, and title fields are text fields and will be searched by default if no search field is specified in a search request. The contents of those fields can also be returned in search results.
 - The genre field is configured as a literal field and is designated as a facet so it can be used to sort and filter the results. Because it's a facet, it cannot be returned in the search results—if you want to retrieve contents of the genre field when you search, you can configure an additional field with the same source data and make it result-enabled. (For more information, see [Configuring Index Fields for an Amazon CloudSearch Domain](#).)
 - The year field is configured as a uint field. You cannot change the configuration of a uint field—uint fields are always search-enabled, facet-enabled, and result-enabled.

When you are finished reviewing the indexing options, click **Continue**.

Create New Search Domain Cancel

NAME YOUR DOMAIN CONFIGURE INDEX **REVIEW INDEX CONFIGURATION** SET UP ACCESS POLICIES CONFIRM

The suggested index configuration is shown below. You can edit these fields or add additional fields. Click **Continue** when you are finished making changes.

Suggested Index Field Configuration

Name	Type	Search	Facet	Result	Default Value	Source	Remove
actor	text	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>		[add]	<input type="checkbox"/>
director	text	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>		[add]	<input type="checkbox"/>
genre	literal	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>		[add]	<input type="checkbox"/>
title	text	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>		[add]	<input type="checkbox"/>
year	uint	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		[add]	<input type="checkbox"/>

- On the **SET UP ACCESS POLICIES** step, click **Recommended rules** and click **Continue**. The recommended rules allow access to the search endpoint from all IP addresses, and restrict access to the document service to the IP address you specify.

Important

If you do not configure access rules for your search domain, you will only be able to interact with the domain through the Amazon CloudSearch console. By default, the document service and search service endpoints are configured to block all IP addresses.

Keep in mind that if you do not have a static IP address, you must re-authorize your computer whenever your IP address changes. If your IP address is assigned dynamically, it is also likely that you're sharing that address with other computers on your network. This means that when you authorize the IP address, all computers that share it will be able to access your search domain's document service endpoint.

Getting Started with Amazon CloudSearch Step 2: Create a Search Domain

Create New Search Domain [Cancel]

NAME YOUR DOMAIN | CONFIGURE INDEX | REVIEW INDEX CONFIGURATION | **SET UP ACCESS POLICIES** | CONFIRM

To authorize or block IP addresses, add one or more access policy rules. You can always access all services through the console, regardless of the rules defined here.

Your IP address appears to be **192.0.2.10**. [change](#)

Set my policy to:

- Recommended rules (Search service: Allow all. Document service: Your IP address only)
- Allow only my IP address access to all services (can be used for testing)
- Allow everyone access to all services (not recommended because anyone can upload documents)
- Deny everyone access to all services (except through the console)
- Copy access policy from another domain

Current Policy:

Effect	Service	IP Ranges	Remove
Allow	Search	0.0.0.0 / 0	[Remove]
Allow	Document	192.0.2.10 / 32	[Remove]

[Add a New Rule](#)

< Back [Cancel] [Continue]

7. On the **CONFIRM** step, review the domain configuration and click **Confirm** to create your domain.

Create New Search Domain [Cancel]

NAME YOUR DOMAIN | CONFIGURE INDEX | REVIEW INDEX CONFIGURATION | SET UP ACCESS POLICIES | **CONFIRM**

Review the information below, then click **Confirm**.

Search Domain Name imdb-movies [Edit Domain Name](#)

Index Fields 5 index fields:

- actor
- director
- genre
- title
- show 1 more...

[Edit Index Fields](#)

Access Policies 2 rules:

- Allow search service for 0.0.0.0/0
- Allow document service for 192.0.2.10/32

[Edit Access Policies](#)

< Back [Cancel] [Confirm]

8. Once the domain has been created, click **OK** to exit the Create New Search Domain wizard and go to the domain's dashboard.

Create New Search Domain [Cancel]

NAME YOUR DOMAIN | CONFIGURE INDEX | REVIEW INDEX CONFIGURATION | SET UP ACCESS POLICIES | **CONFIRM**

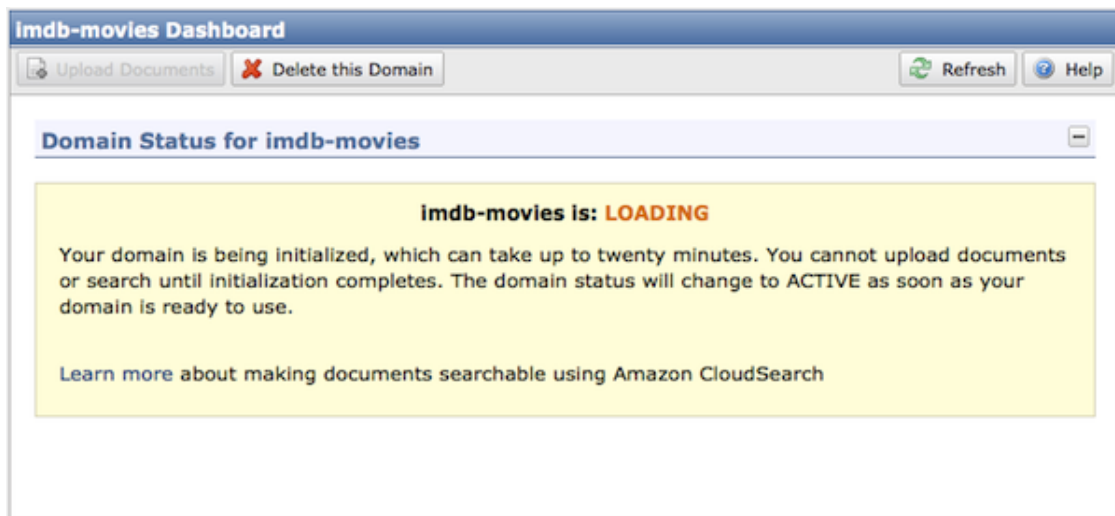
Congratulations

You have successfully created the domain **movies**. The domain will take several minutes to initialize. Once the domain is initialized, you may upload documents, and make any other changes to the domain.

[OK]

When you create a new domain, Amazon CloudSearch initializes resources for the domain, which can take around half an hour. During this initialization process, the status of the domain will be **LOADING**.

You can begin uploading the data you want to search as soon as the domain status changes to PROCESSING. Once the status changes to ACTIVE, your domain will be fully-functional and available to process search requests.



Note

While you can start uploading documents through the console once the domain status reaches the PROCESSING state, you won't be able to upload data through the command line tools or document service API until the domain status is ACTIVE.

Step 3: Send Data to Amazon CloudSearch for Indexing

You upload the data you want to search to your domain so that Amazon CloudSearch can build and deploy a searchable index. The format used to submit documents to Amazon CloudSearch is called Search Data Format (SDF). The AWS Management Console can automatically generate SDF from several types of files:

- Comma Separated Value (.csv)
- Adobe Portable Document Format (.pdf)
- HTML (.htm, .html)
- Microsoft Excel (.xls, .xlsx)
- Microsoft PowerPoint (.ppt, .pptx)
- Microsoft Word (.doc, .docx)
- Text Documents (.txt)
- JSON Documents (.json)
- XML Documents (.xml)

For most file types, including JSON and XML, Amazon CloudSearch adds a single add document operation to the SDF batch for each source file. If metadata is available for the file, the metadata is mapped to corresponding document fields—the fields generated from the document metadata vary depending on the file type. The contents of the source file are parsed into a single text field. If the file contains more than 1 MB of data, the data mapped to the text field is truncated so that the document does not exceed 1 MB.

Getting Started with Amazon CloudSearch

Step 3: Send Data for Indexing

CSV files are handled differently. When you upload a CSV file, Amazon CloudSearch uses the contents of the first row to define the document fields, and creates a separate document each following row. If there is a column header called *docid*, the values in that column are used as the document IDs. If necessary, the docid values are normalized to conform to the allowed character set: a-z (lower-case letters), 0-9, and _ (underscore). If there is no docid column, a unique ID is generated for each document based on the filename and row number. Similarly, if there is a column called *version*, the values in that column are used as the versions for the document updates. Version numbers must be specified as 32-bit unsigned integers. If there is no version column, version numbers are generated based on a timestamp.

If you upload multiple types of files, CSV files are parsed row-by-row, and non-CSV files are treated as individual documents.

The sample IMDB movies data is already formatted as SDF and contains add requests for over 5,000 popular movies. Each add request specifies a unique ID for the movie, a document version number, and fields that contain the movie data such as title and genre.

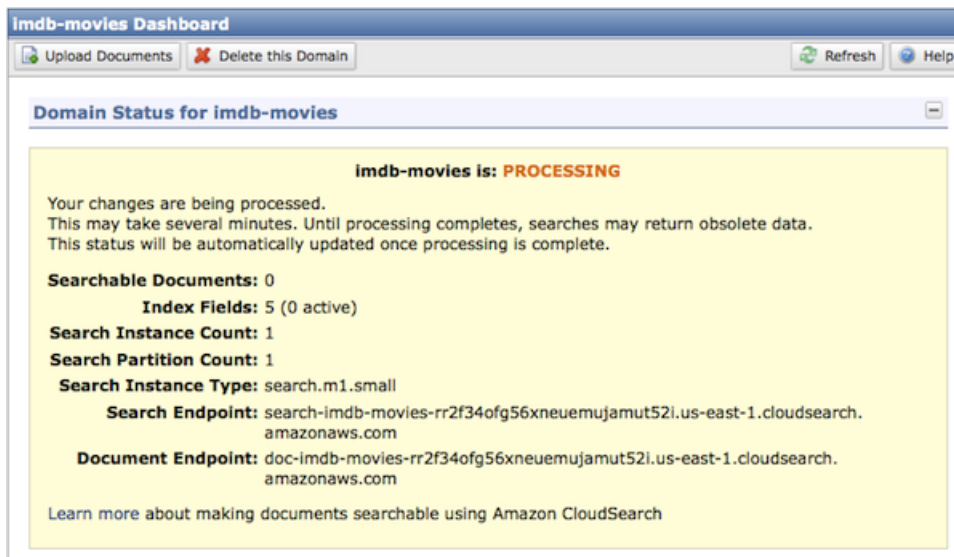
This tutorial shows how to submit data through the Amazon CloudSearch console, but you can also [generate SDF](#) and [upload data](#) with the command line tools, and submit SDF batches through the [DocumentsBatch API](#).

To add the sample data to your movies domain

1. Go to the Amazon CloudSearch console at [Amazon CloudSearch console](#).
2. In the **Navigation** panel, click the name of your movies domain to view the domain dashboard.
3. At the top of the domain dashboard, click the **Upload Documents** button.

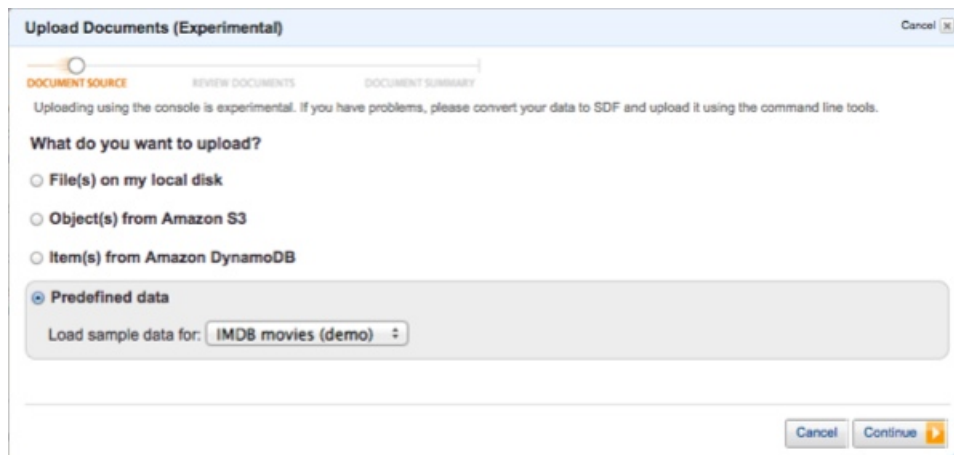
Note

The **Upload Documents** button is available once the domain status is PROCESSING or ACTIVE. You will not be able to search uploaded documents until the domain status is ACTIVE.



4. On the **DOCUMENT SOURCE** step, select **Predefined data**, choose **IMDB movies (demo)**, and click **Continue**.

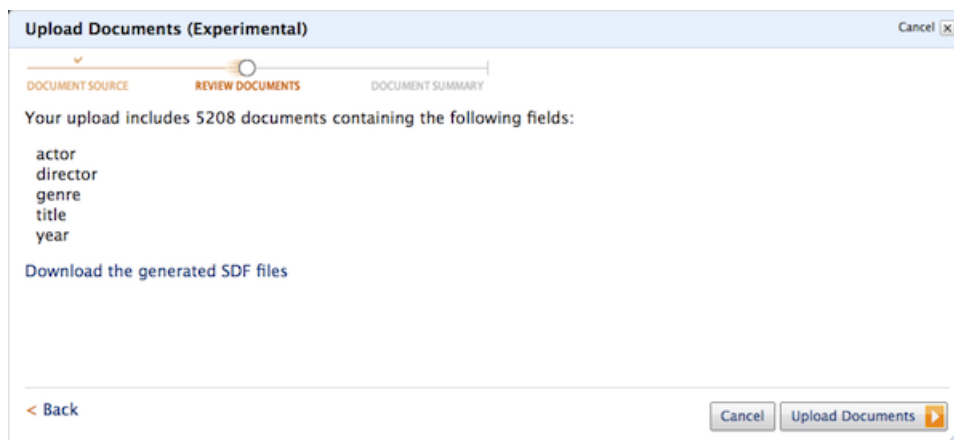
Getting Started with Amazon CloudSearch Step 3: Send Data for Indexing



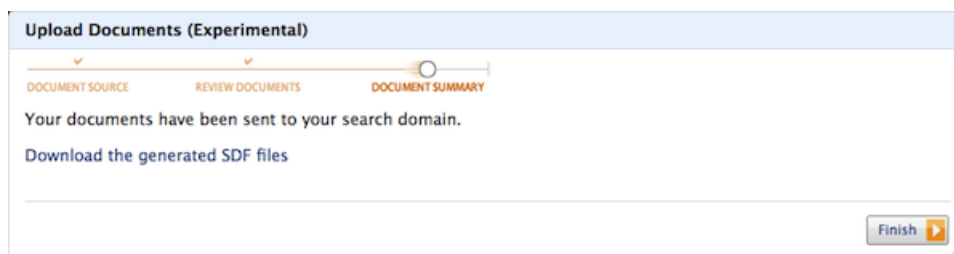
5. On the **REVIEW DOCUMENTS** step, review the upload summary and click **Upload Documents** to send the data to your domain for indexing.

Note

If you'd like to see what the SDF data looks like, click **Download the generated SDF files**. For more information about SDF and preparing your own data, see [Preparing Your Data for Amazon CloudSearch](#).



6. On the **DOCUMENT SUMMARY** step, click **Finish** to return to the domain dashboard.



That's it! You now have a fully functional Amazon CloudSearch domain that you can start searching. The data is automatically indexed in near real-time, so you can start searching your domain right away.

Step 4: Search Your Amazon CloudSearch Domain

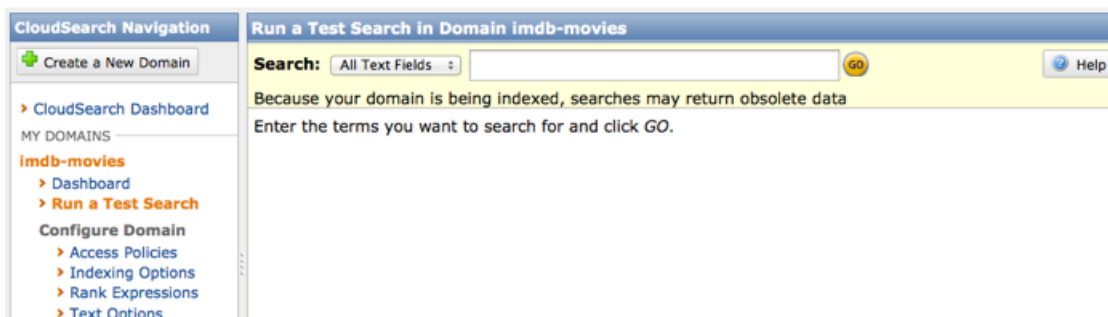
You can use the search tester in the Amazon CloudSearch console to perform simple text searches. To perform more complex Boolean queries, you can submit search requests through a Web browser or send HTTP requests using cURL or any HTTP library.

Searching with the Search Tester

The search tester enables you to choose which fields you want to search, sort the results, and browse any facets that are configured for the domain. By default, results are sorted according to an automatically-generated relevance score, *text_relevance*. (For more information about customizing how results are ranked, see [Customizing Result Ranking with Amazon CloudSearch](#).)

To search your domain

1. Go to the Amazon CloudSearch console at [Amazon CloudSearch console](#).
2. In the **Navigation** panel, click the name of your movies domain.
3. In the **Navigation** panel, click the **Run a Test Search** link for your movies domain.



4. Select the field(s) you want to search, enter the text you want to search for, and click **Go**.



To view the HTTP search request that was sent to your domain's search endpoint and the JSON or XML response returned by Amazon CloudSearch, click the **view raw** link for the response format you want to see.

You can copy and paste the request URL to submit the request and view the response from a Web browser. Requests can be sent via HTTP or HTTPS.

Submitting Search Requests from a Web Browser

To perform more complex searches, you can submit your own search requests directly to your search endpoint. You can perform simple and Boolean searches and specify a variety of options to constrain your search, request facet information, customize ranking, and control what information is returned in the results.

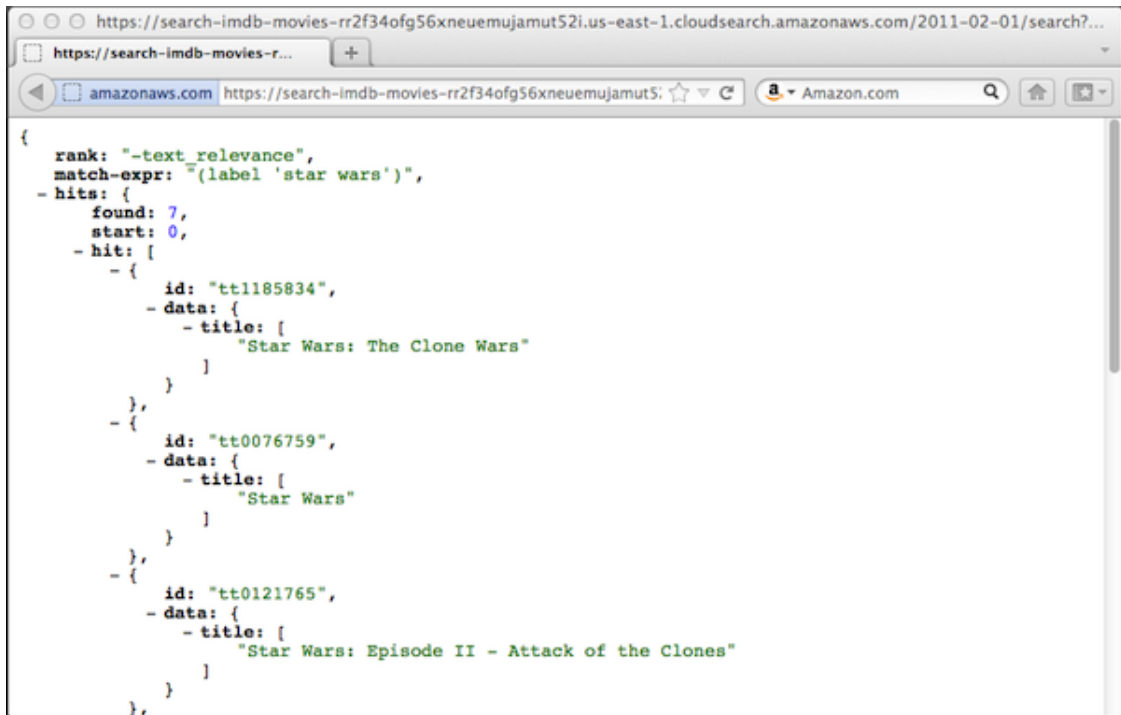
For example, to search your movies domain and get the titles of all of the available *Star Wars* movies, append the following search string to your search endpoint. (2011-02-01 is the API version and must be specified.)

```
/2011-02-01/search?q=star+wars&return-fields=title
```

Note

Your domain's search endpoint is shown on the domain dashboard. You can also perform a search from the AWS Management Console, view the raw request and response, and copy the request URL from the Search Request field.

By default, Amazon CloudSearch returns the response in JSON. You can also get the search results formatted in XML by specifying the results-type parameter, `results-type=xml`. (Errors are always returned in JSON.) The following image shows the results of the previous query.



Filtering Results

You can use the Boolean query option, `bq`, to find documents that have particular numeric attributes. You can filter based on an exact value in a field, an inequality, or a range of values, as in these examples:

- `bq=year:2000` matches documents with the year 2000.

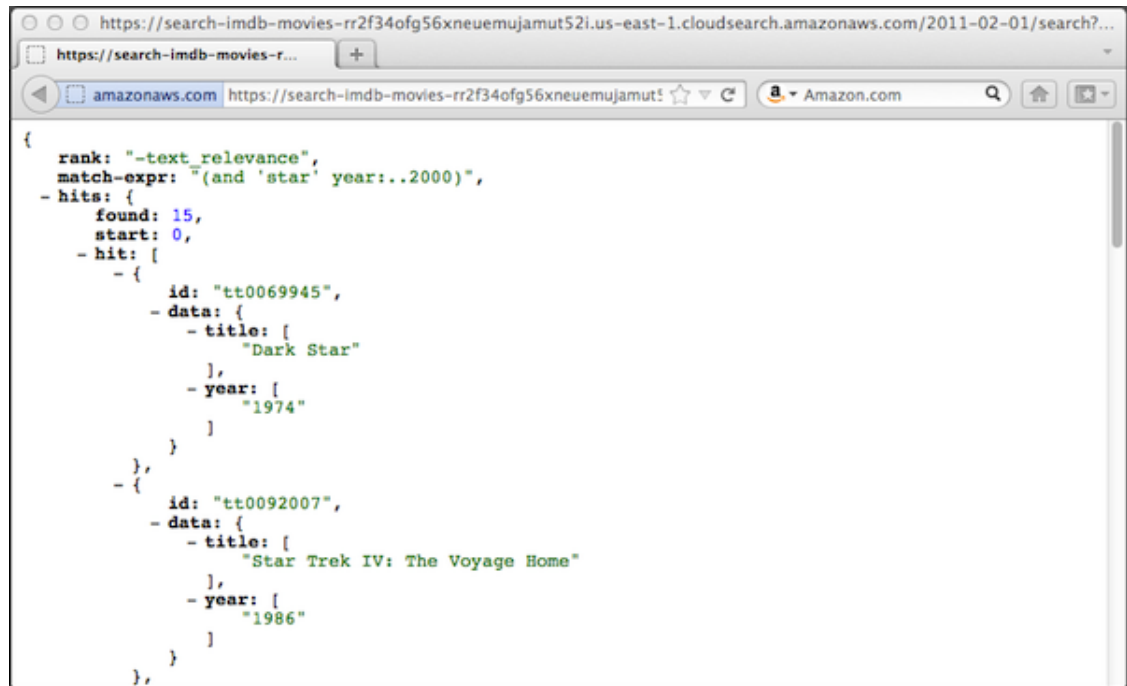
Getting Started with Amazon CloudSearch Submitting Search Requests from a Web Browser

- `bq=year:2000..` matches documents with a year greater than or equal to 2000
- `bq=year:..2000` matches documents with a year less than or equal to 2000
- `bq=year:2000..2011` matches documents with a year between 2000 and 2011, inclusive.

For example, the following Boolean query searches for "star", finds all of the matching movies that were released before 2000, and returns title and year of each one:

```
2011-02-01/search?bq=(and 'star' year:..2000)&return-fields=title,year
```

The response shows the number of matching documents and the requested fields for each hit.



```
{
  rank: "-text_relevance",
  match-expr: "(and 'star' year:..2000)",
  - hits: {
    found: 15,
    start: 0,
    - hit: {
      - {
        id: "tt0069945",
        - data: {
          - title: [
            "Dark Star"
          ],
          - year: [
            "1974"
          ]
        }
      },
      - {
        id: "tt0092007",
        - data: {
          - title: [
            "Star Trek IV: The Voyage Home"
          ],
          - year: [
            "1986"
          ]
        }
      }
    }
  },
}
```

For more information about constructing search queries, see [Searching Your Data with Amazon CloudSearch](#).

Ranking the Search Results

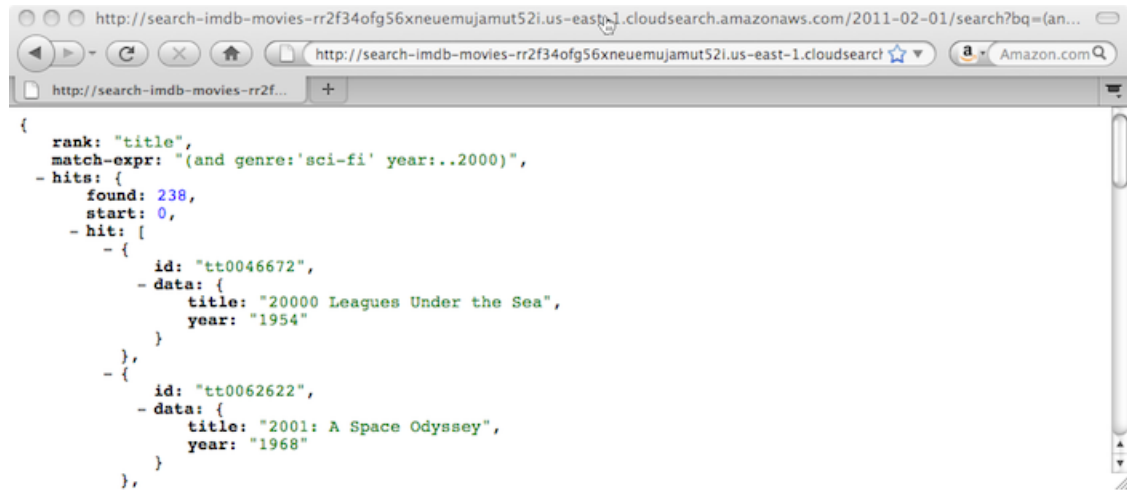
By default, Amazon CloudSearch ranks the search results according to an automatically generated `text_relevance` score. You can change how results are ranked by specifying the `rank` option in your search request to specify the field or rank expression you want to use for ranking. (A rank expression is a custom numeric expression that can be evaluated for each document in the set of matching documents. For information about defining your own rank expressions, see [Customizing Result Ranking with Amazon CloudSearch](#).)

If you specify a text field with the rank option, the results are sorted alphabetically according to that field. For example, to rank results from your movies domain alphabetically by title, add `&rank=title` to your query string:

```
2011-02-01/search?bq=(and genre:'sci-fi' year:..2000)&return-
fields=title,year&rank=title
```

Getting Started with Amazon CloudSearch Submitting Search Requests from a Web Browser

When you rank alphabetically, the results are sorted in ascending order by default. Any values that begin with a numeral are listed before the first *A* entry:



```
{
  rank: "title",
  match-expr: "(and genre:'sci-fi' year:..2000)",
  - hits: {
    found: 238,
    start: 0,
    - hit: [
      - {
        id: "tt0046672",
        - data: {
          title: "20000 Leagues Under the Sea",
          year: "1954"
        }
      },
      - {
        id: "tt0062622",
        - data: {
          title: "2001: A Space Odyssey",
          year: "1968"
        }
      }
    ]
  }
}
```

Similarly, you can specify an integer field with the `rank` option to sort the results numerically.

By default, when you rank alphabetically or numerically, results are returned in ascending order. You can prefix the field name with a minus (-) if you want the results returned in descending order. If you specify a comma separated list of fields or rank expressions, the first field or expression is used as the primary sort criteria, the second is used as the secondary sort criteria, and so on.

For more information about ranking results, see [Customizing Result Ranking with Amazon CloudSearch](#)

Getting Facet Information

A facet is an index field that represents a category that you want to use to refine and filter search results. When you submit search requests to Amazon CloudSearch, you can request facet information to find out how many hits share the same value in a facet. You can display this information along with the search results and use it to enable users to interactively refine their searches. (This is often referred to as faceted navigation or faceted search.)

A facet can be any numeric field or a text or literal field that has faceting enabled in your domain configuration. To request facet information in your search request, you specify:

- One or more facets
- Facet constraints that specify the particular values you want to count (optional)
- How you want the facet values to be sorted in the results (optional)

For each facet, Amazon CloudSearch calculates the number of hits that share the same value. If you specify constraints, the facet counts are calculated only for values that match the constraints. Only constraints that have matches are included in the facet results.

Note

Values from a facet-enabled text or literal field cannot be returned in the search results. Text and literal fields can be facet-enabled or result-enabled, but not both. If you want to return the value from an SDF document field as well as use the field as a facet, create two index fields that use the same SDF document field as a source and make one result-enabled, and the other facet-enabled.

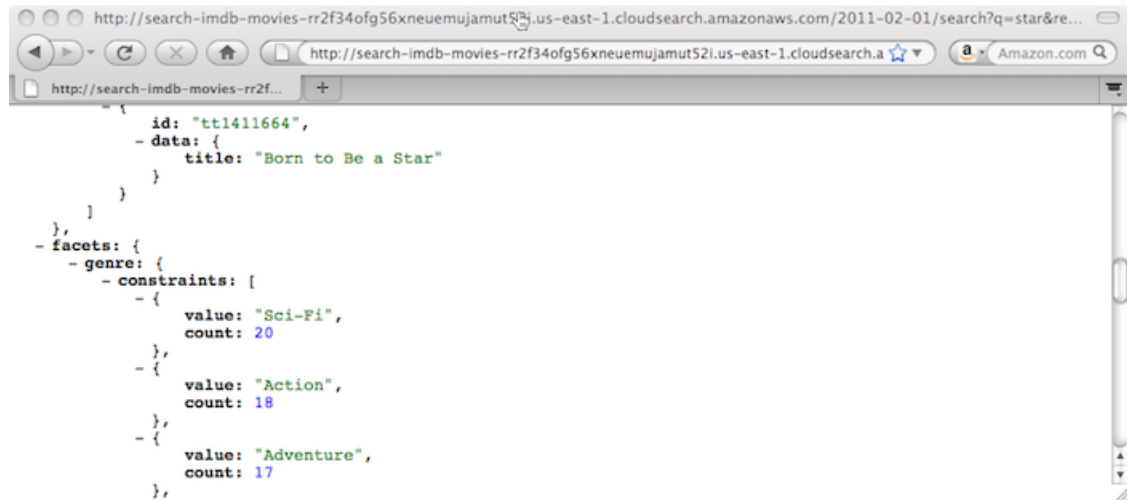
Getting Started with Amazon CloudSearch Submitting Search Requests from a Web Browser

To get facet counts with your search results

- Use the *facet* option to specify the fields for which you want to compute facets. For the sample IMDB movies data faceting is enabled for one field, *genre*.

```
/2011-02-01/search?q=star&return-fields=title&facet=genre
```

The facets appear below the hits in the results.

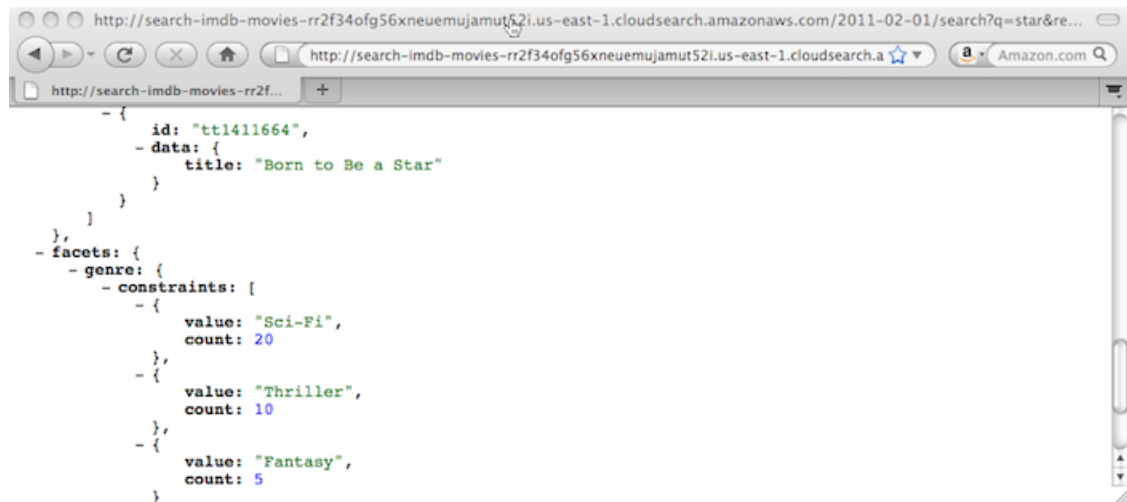


```
    {
      id: "tt1411664",
      - data: {
        title: "Born to Be a Star"
      }
    },
  ],
  - facets: {
    - genre: {
      - constraints: [
        - {
          value: "Sci-Fi",
          count: 20
        },
        - {
          value: "Action",
          count: 18
        },
        - {
          value: "Adventure",
          count: 17
        }
      ],
    }
  }
}
```

If you want to compute facet counts for selected values of a facet field, you can set facet constraints for the field. Facet constraints do not constrain the results themselves, only the facet counts that are returned. For example, the following request only counts the movies that are in the Sci-Fi, Fantasy, or Thriller genres:

```
/2011-02-01/search?q=star&return-fields=title&facet=genre&facet-genre-constraints='Sci-Fi', 'Fantasy', 'Thriller'
```

This constrains the facet counts to the three specified values:



```
    {
      id: "tt1411664",
      - data: {
        title: "Born to Be a Star"
      }
    },
  ],
  - facets: {
    - genre: {
      - constraints: [
        - {
          value: "Sci-Fi",
          count: 20
        },
        - {
          value: "Thriller",
          count: 10
        },
        - {
          value: "Fantasy",
          count: 5
        }
      ],
    }
  }
}
```


Getting Started with Amazon CloudSearch Step 5: Delete Your Amazon CloudSearch Movies Domain

For more information about faceted searches, see [Getting and Using Facet Information in Amazon CloudSearch](#).

Step 5: Delete Your Amazon CloudSearch Movies Domain

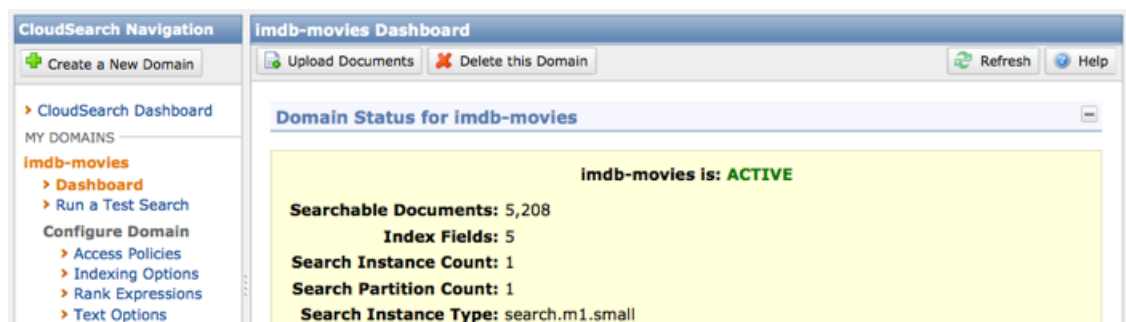
When you are finished experimenting with your movies domain, you need to delete it to avoid incurring additional usage fees.

Important

Deleting a domain deletes the index associated with the domain and takes the domain's document and search endpoints offline permanently.

To delete your imdb-movies domain

1. Go to the Amazon CloudSearch console at [Amazon CloudSearch console](#).
2. In the **Navigation** panel, click the name of your movies domain to view to the domain dashboard.
3. At the top of the domain dashboard, click the **Delete this Domain** button.



4. In the **Delete Domain** dialog box, select the **Delete the domain** option and click **OK** to permanently remove the domain and all of its data.



Note

It can take around 15 minutes to delete the domain and its resources. Until then, the domain status will be *BEING DELETED*.

Wondering where to go next? [What is Amazon CloudSearch](#) has a guide to the rest of the Amazon CloudSearch developer documentation. For more information about the Amazon CloudSearch query language, see [Searching Your Data with Amazon CloudSearch](#). If you're ready to set up a domain with your own data, see [Preparing Your Data for Amazon CloudSearch](#) and [Uploading Data to an Amazon CloudSearch Domain](#) for information about formatting and submitting your data to Amazon CloudSearch.