GitHub

PDF (/pdfs/kendra/latest/dg/kendra-dg.pdf#data-source-github) | **RSS (amazon-kendra-release-notes.rss)**

GitHub is a web-based hosting service for software development providing code storage and management services with version control. You can use Amazon Kendra to index your GitHub Enterprise Cloud (SaaS) and GitHub Enterprise Server (On Prem) repository files, issue and pull requests, issue and pull request comments, and issue and pull request comment attachments. You can also choose to include or exclude certain files.

You can connect Amazon Kendra to your GitHub data source using the Amazon Kendra console (https://console.aws.amazon.com/kendra/) and the GitHubConfiguration (https://docs.aws.amazon.com/kendra/latest/APIReference/API_GitHubConfiguration.html) API.

For troubleshooting your Amazon Kendra GitHub data source connector, see Troubleshooting data sources (./troubleshooting-data-sources.html) .

Topics

Supported features

Amazon Kendra GitHub data source connector supports the following features:

- Change log
- Field mappings
- User context filtering
- Inclusion/exclusion filters

Prerequisites

Before you can use Amazon Kendra to index your GitHub data source, make these changes in your GitHub and AWS accounts.

In GitHub, make sure you have:

- Created a GitHub user with administrative permissions to the GitHub organization.
- Created a personal access token for authentication credentials. See GitHub documentation on creating a personal access token (https://docs.github.com/en/authentication/keeping-your-account-and-data-secure/creating-a-personal-access-token) .
- **Recommended:**Created an OAuth token for authentication credentials. Use OAuth token for better API throttle limits and connector performance. See GitHub documentation on OAuth authorization (https://docs.github.com/en/rest/apps/oauth-applications?apiVersion=2022-11-28#about-oauth-apps-and-oauth-authorizations-of-github-apps) .
- **Optional:** Installed a SSL certificate.
- Noted the GitHub host URL for the type of GitHub service that you use. For example, the host URL for GitHub cloud could be *https://api.github.com* and the host URL for GitHub server could be *https://on-prem-host-url/api/v3/*.
- Noted the GitHub organization name for your respositories from your GitHub settings.
- Added the following permissions:

  For GitHub Enterprise Cloud (SaaS)

  - repo:status
  - public_repo
  - repo:invite
  - read:org
  - user:email
  - read:user

  For GitHub Enterprise Server (On Prem)

  - repo:status
  - public_repo
  - repo:invite
  - read:org
  - user:email
  - read:user
  - site_admin

- Checked each document is unique in GitHub and across other data sources you plan to use for the same index. Each data source that you want to use for an index must not contain the same document across the data sources. Document IDs are global to an index and must be unique per index.

In your AWS account, make sure you have:

- Created an Amazon Kendra index (https://docs.aws.amazon.com/kendra/latest/dg/create-index.html) and, if using the API, noted the index ID.
- Created an IAM role (https://docs.aws.amazon.com/kendra/latest/dg/iam-roles.html#iam-roles-ds) for your data source and, if using the API, noted the ARN of the IAM role.

> **Note**
>
> If you change your authentication type and credentials, you must update your IAM role to access the correct AWS Secrets Manager secret ID.

- Stored your GitHub authentication credentials in an AWS Secrets Manager secret and, if using the API, noted the ARN of the secret.

> **Note**
>
> It's recommended that you regularly refresh or rotate your credentials and secret. Provide only the necessary access level for your own security. It's **not** recommended to re-use credentials and secrets across data sources, and connector versions 1.0 and 2.0 (where applicable).

If you don't have an existing IAM role or secret, you can use the console to create a new IAM role and Secrets Manager secret when you connect your GitHub data source to Amazon Kendra. If you are using the API, you must provide the ARN of an existing IAM role and Secrets Manager secret, and an index ID.

Connection instructions

To connect Amazon Kendra to your GitHub data source, you must provide the necessary details of your GitHub data source so that Amazon Kendra can access your data. If you have not yet configured GitHub for Amazon Kendra, see Prerequisites (#prerequisites-github) .

Console API

To connect Amazon Kendra to GitHub

You must specify the following using the GitHubConfiguration (https://docs.aws.amazon.com/kendra/latest/APIReference/API_GitHubConfiguration.html) API object:

- **Data source type**—Specify the data source type as either `SAAS` or `ON_PREMISE`.
- **Secret Amazon Resource Name (ARN)**—Provide the Amazon Resource Name (ARN) of a Secrets Manager secret that contains the authentication credentials for your GitHub account. The secret is stored in a JSON structure with the following keys:

```
{
    "personalToken": "token"
}
```

> **Note**
>
> It's recommended that you regularly refresh or rotate your credentials and secret. Provide only the necessary access level for your own security. It's **not** recommended to re-use credentials and secrets across data sources, and connector versions 1.0 and 2.0 (where applicable).

- **IAM role**—Specify `RoleArn` when you call `CreateDataSource` to provide an IAM role with permissions to access your Secrets Manager secret and to call the required public APIs for the GitHub connector and Amazon Kendra. For more information, see IAM roles for GitHub data sources (https://docs.aws.amazon.com/kendra/latest/dg/iam-roles.html#iam-roles-ds) .

You can also add the following optional features:

- **Virtual Private Cloud (VPC)**—Specify `VpcConfiguration` as part of the data source configuration. See Configuring Amazon Kendra to use a VPC (https://docs.aws.amazon.com/kendra/latest/dg/vpc-configuration.html) .

> **Note**
>
> If you use GitHub server, you must use an Amazon VPC to connect to your GitHub server.

- **Change log**—Whether Amazon Kendra should use the GitHub data source change log mechanism to determine if a document must be added, updated, or deleted in the index.

> **Note**
>
> Use the change log if you don't want Amazon Kendra to scan all of the documents. If your change log is large, it might take Amazon Kendra less time to scan the documents in the GitHub data source than to process the change log. If you are syncing your GitHub data source with your index for the first time, all documents are scanned.

- **Inclusion and exclusion filters**—Specify whether to include or exclude certain files.

> **Note**
>
> Most data sources use regular expression patterns, which are inclusion or exclusion patterns referred to as filters. If you specify an inclusion filter, only content that matches the inclusion filter is indexed. Any document that doesn't match the inclusion filter isn't indexed. If you specify an inclusion and exclusion filter, documents that match the exclusion filter are not indexed, even if they match the inclusion filter.

- **Field mappings**—Choose to map your GitHub data source fields to your Amazon Kendra index fields. For more information, see Mapping data source fields (https://docs.aws.amazon.com/kendra/latest/dg/field-mapping.html) .
- **User context filtering**—Amazon Kendra crawls the access control list (ACL) for your data source by default. The ACL information is used to filter search results based on the user or their group access to documents. For more information, see User context filtering for GitHub data sources (https://docs.aws.amazon.com/kendra/latest/dg/user-context-filter.html#datasource-context-filter) .

Learn more

To learn more about integrating Amazon Kendra with your GitHub data source, see:

- Reimagine search on GitHub repositories with the power of the Amazon Kendra GitHub connector (https://aws.amazon.com/blogs/machine-learning/reimagine-search-on-github-repositories-with-the-power-of-the-amazon-kendra-github-connector/)