Cloudera DataFlow

Slack to S3/ADLS

Date published: 2021-04-06
Date modified: 2024-01-09

https://docs.cloudera.com/

Legal Notice

The documentation is and contains Cloudera proprietary information protected by copyright and other intellectual property rights. No license under copyright or any other intellectual property right is granted herein.

Unless otherwise noted, scripts and sample code are licensed under the Apache License, Version 2.0.

Copyright information for Cloudera software may be found within the documentation accompanying each component in a particular release.

Cloudera software includes software from various open source or other third party projects, and may be released under the Apache Software License 2.0 (“ASLv2”), the Affero General Public License version 3 (AGPLv3), or other license terms. Other software included may be released under the terms of alternative open source licenses. Please review the license and notice files accompanying the software for additional licensing information.

Please visit the Cloudera software product page for more information on Cloudera software. For more information on Cloudera support services, please visit either the Support or Sales page. Feel free to contact us directly to discuss your specific needs.

Cloudera reserves the right to change any products at any time, and without notice. Cloudera assumes no responsibility nor liability arising from the use of products, except as expressly agreed to in writing by Cloudera.

Cloudera, Cloudera Altus, HUE, Impala, Cloudera Impala, and other Cloudera marks are registered or unregistered trademarks in the United States and other countries. All other trademarks are the property of their respective owners.

Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES NOT WARRANT THAT CLOUDERA PRODUCTS NOR SOFTWARE WILL OPERATE UNINTERRUPTED NOR THAT IT WILL BE FREE FROM DEFECTS NOR ERRORS, THAT IT WILL PROTECT YOUR DATA FROM LOSS, CORRUPTION NOR UNAVAILABILITY, NOR THAT IT WILL MEET ALL OF CUSTOMER’S BUSINESS REQUIREMENTS. WITHOUT LIMITING THE FOREGOING, AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, CLOUDERA EXPRESSLY DISCLAIMS ANY AND ALL IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, QUALITY, NON-INFRINGEMENT, TITLE, AND FITNESS FOR A PARTICULAR PURPOSE AND ANY REPRESENTATION, WARRANTY, OR COVENANT BASED ON COURSE OF DEALING OR USAGE IN TRADE.

Contents

ReadyFlow overview: Slack to S3/ADLS

You can use the Slack to S3/ADLS ReadyFlow to consume events from Slack, convert them to Avro, CSV, or JSON format and write them to a CDP managed destination in Amazon S3 or Azure Data Lake Service (ADLS).

This ReadyFlow consumes events from a Slack App, converts them to the specified output data format, and writes them to a CDP managed destination S3 or ADLS location. For the source, subscribe to the events to be notified of in Slack. For the destination, specify the S3 or ADLS storage location and path. The flow writes out a file every time its size has either reached 100MB or five minutes have passed. Files can reach a maximum size of 1GB. Failed S3 or ADLS write operations are retried automatically to handle transient issues. Define a KPI on the failure_WriteToS3/ADLS connection to monitor failed write operations.

Note: This ReadyFlow leverages CDP's centralized access control for cloud storage access. Make sure to either set up Ranger policies or an IDBroker mapping allowing your workload user access to the target S3 or ADLS location.

Slack to S3/ADLS ReadyFlow details	
Source	Slack
Source Format	Slack
Destination	CDP managed Amazon S3 or ADLS
Destination Format	Avro, CSV, or JSON

Prerequisites

Learn how to collect the information you need to deploy the Slack to S3/ADLS ReadyFlow, and meet other prerequisites.

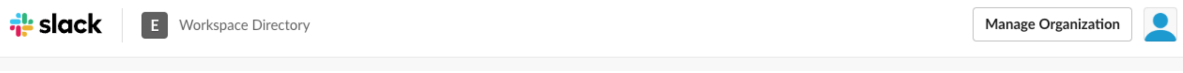
For your data ingest source

- You have a Slack sandbox account.

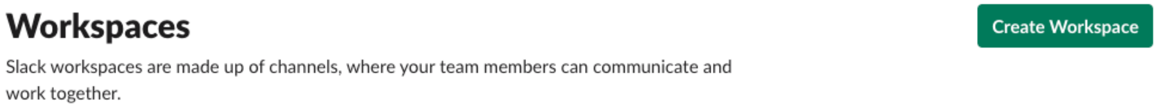
  If you already have a Slack environment, you may skip this step. However you are highly recommended to make changes in dev/test environments before moving them to production. Slack offers a sandbox environment where such experiments can be carried out.

- You have created a Slack workspace.

1. In Slack Workspace Directory click Manage Organization.



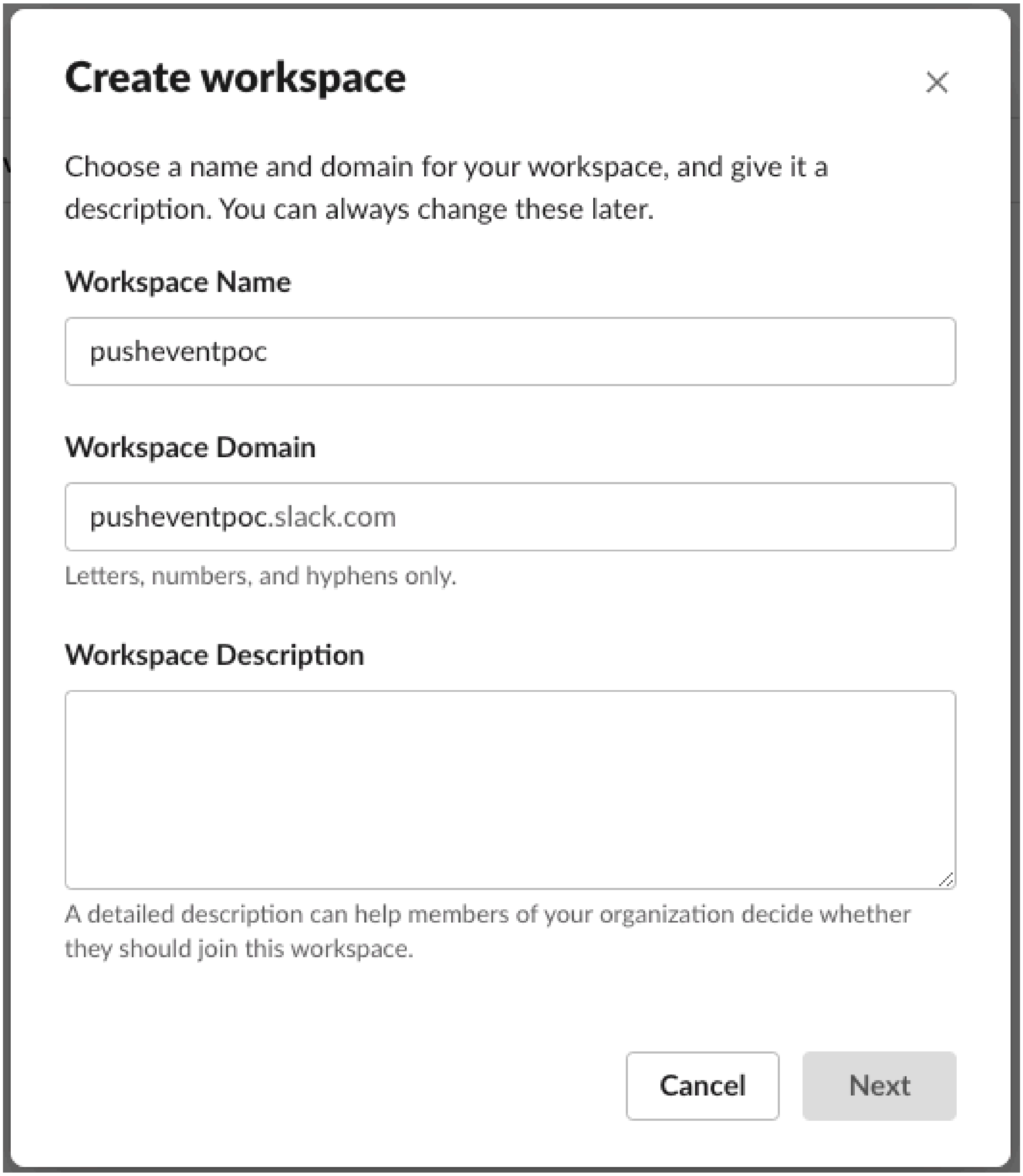
2. Click Create Workspace.



Workspaces

Slack workspaces are made up of channels, where your team members can communicate and work together.

3. Provide Workspace Name and Workspace Domain. You may leave Workspace Description empty.



Create workspace

Choose a name and domain for your workspace, and give it a description. You can always change these later.

Workspace Name

pusheventpoc

Workspace Domain

pusheventpoc.slack.com

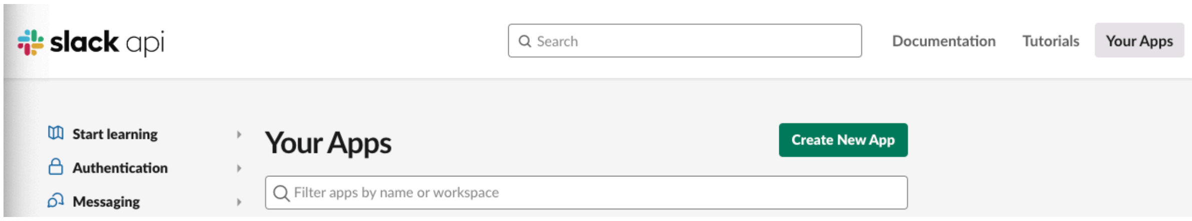
Letters, numbers, and hyphens only.

Workspace Description

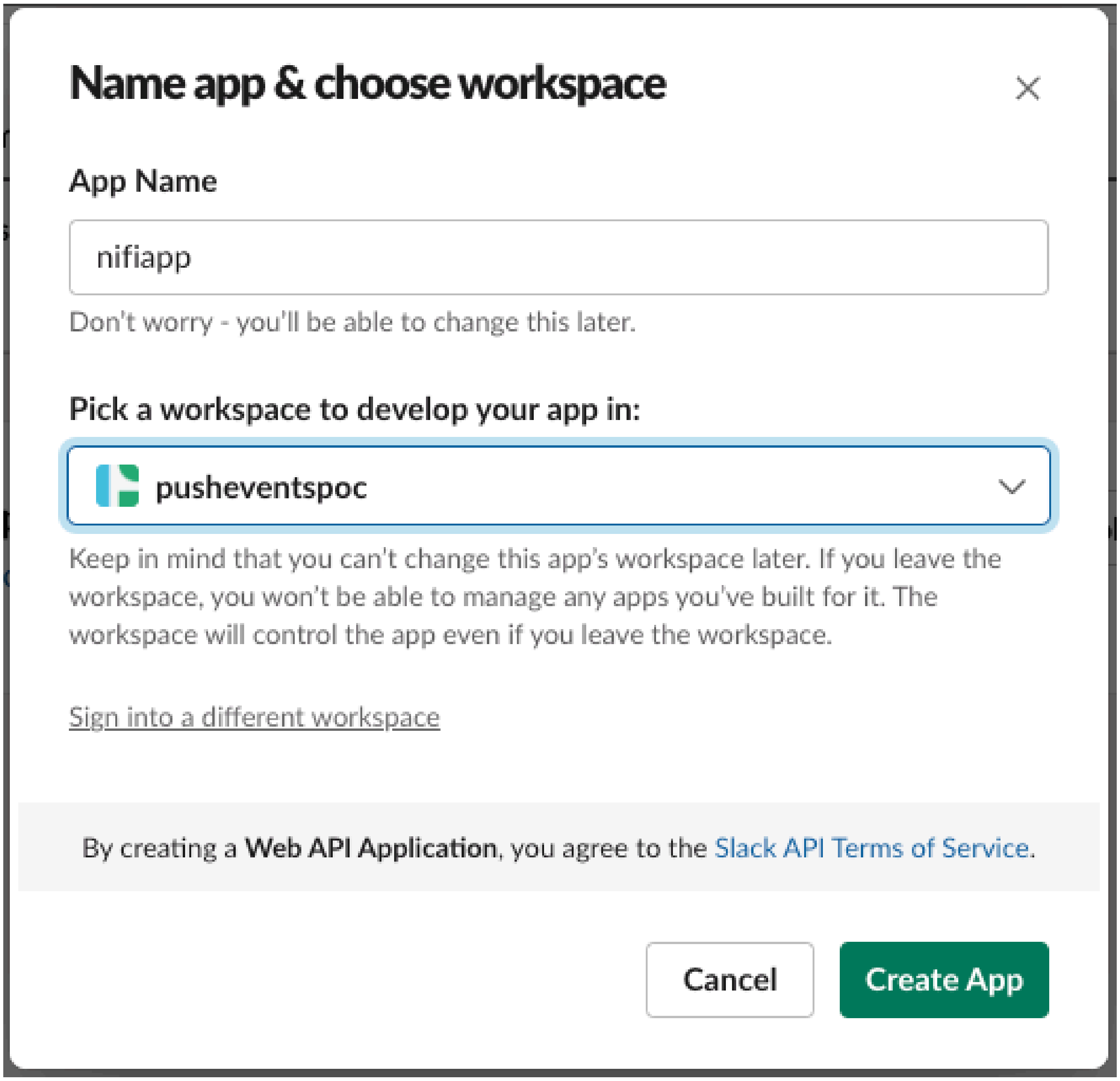
A detailed description can help members of your organization decide whether they should join this workspace.

- You have created, configured, and installed a Slack App.

1. Go to https://api.slack.com/apps, and click on Create New App.



2. Select From Scratch.
3. Provide an App Name.
4. Under Pick a workspace to develop your app in: select the workspace you created in the previous step.



Name app & choose workspace

App Name

nifiapp

Don't worry - you'll be able to change this later.

Pick a workspace to develop your app in:

pusheventspoc

Keep in mind that you can't change this app's workspace later. If you leave the workspace, you won't be able to manage any apps you've built for it. The workspace will control the app even if you leave the workspace.

Sign into a different workspace

By creating a **Web API Application**, you agree to the Slack API Terms of Service.

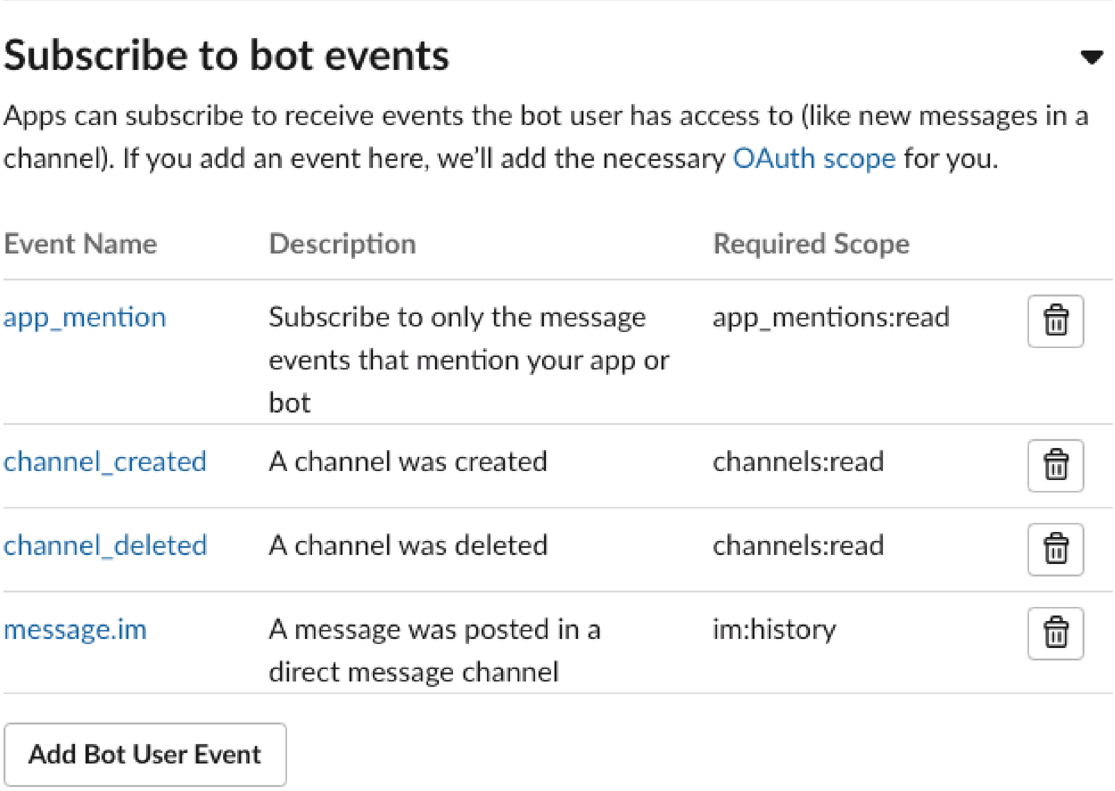
5. Go to https://api.slack.com/apps/ and select the application you have created.
6. Select Event Subscription and click on the toggle to enable events.

   After enabling event subscription an input box appears.
7. Provide a request URL.

   The request URL consists of the DataFlow Inbound Connection Endpoint Hostname, the value of the Allowed Paths parameter (if specified), and the Listening Port in the following format https://[****Endpoint*

Hostname***]:[***Listening Port***][***/Allowed Path***]. For example, https://cdev-test.inbound.dfx.nnfjwxkq.xcu2-8y8x.dev.cldr.work:9876/events

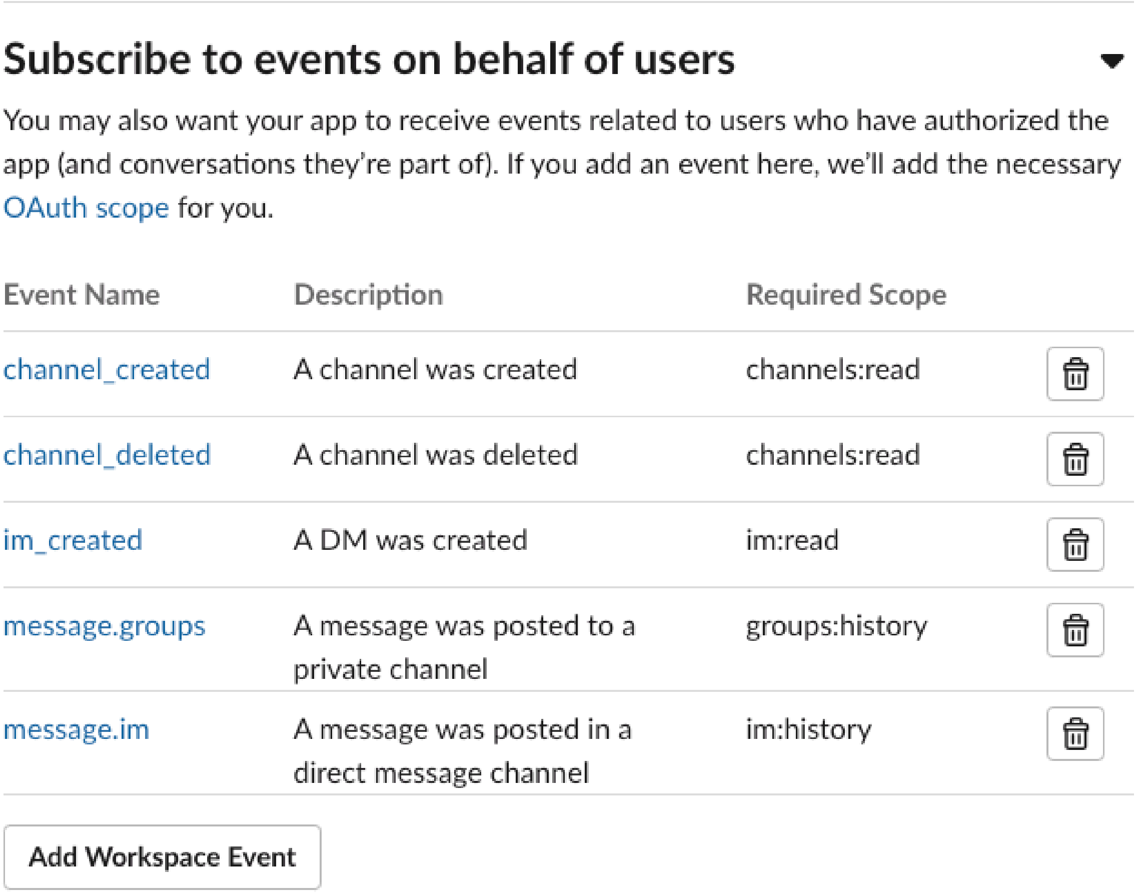
8. Select the events you want to subscribe to. Only the events configured here will be pushed to CDF.



Subscribe to bot events

Apps can subscribe to receive events the bot user has access to (like new messages in a channel). If you add an event here, we'll add the necessary OAuth scope for you.

Event Name	Description	Required Scope	
app_mention	Subscribe to only the message events that mention your app or bot	app_mentions:read	🗑
channel_created	A channel was created	channels:read	🗑
channel_deleted	A channel was deleted	channels:read	🗑
message.im	A message was posted in a direct message channel	im:history	🗑

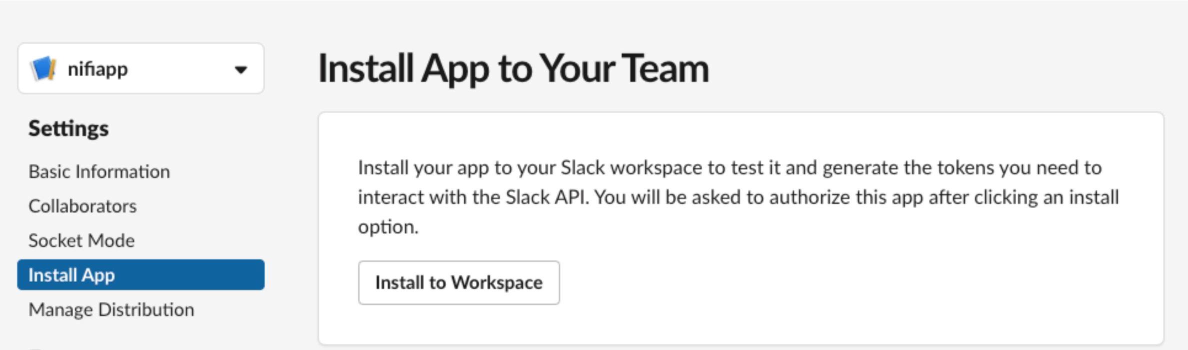


Subscribe to events on behalf of users

You may also want your app to receive events related to users who have authorized the app (and conversations they're part of). If you add an event here, we'll add the necessary OAuth scope for you.

Event Name	Description	Required Scope
channel_created	A channel was created	channels:read
channel_deleted	A channel was deleted	channels:read
im_created	A DM was created	im:read
message.groups	A message was posted to a private channel	groups:history
message.im	A message was posted in a direct message channel	im:history

9. Install the Slack App by clicking Install App in the left pane then Install to Workspace in the main pane.



Install App to Your Team

Install your app to your Slack workspace to test it and generate the tokens you need to interact with the Slack API. You will be asked to authorize this app after clicking an install option.

For more information on getting started with Slack Enterprise Grid sandboxes, see the Slack documentation.

For DataFlow

- Your environment has public subnets and public connectivity enabled.
- You have enabled DataFlow for an environment.

  For information on how to enable DataFlow for an environment, see Enabling DataFlow for an Environment.
- You have created a Machine User to use as the CDP Workload User.

- You have given the CDP Workload User the EnvironmentUser role.

  1. From the Management Console, go to the environment for which DataFlow is enabled.
  2. From the Actions drop down, click Manage Access.
  3. Identify the user you want to use as a Workload User.

     Note:

     The CDP Workload User can be a machine user or your own user name. It is best practice to create a dedicated Machine user for this.
  4. Give that user EnvironmentUser role.
- You have synchronized your user to the CDP Public Cloud environment that you enabled for DataFlow.

  For information on how to synchronize your user to FreeIPA, see Performing User Sync.
- You have granted your CDP user the DFCatalogAdmin and DFFlowAdmin roles to enable your user to add the ReadyFlow to the Catalog and deploy the flow definition.

  1. Give a user permission to add the ReadyFlow to the Catalog.

     a. From the Management Console, click User Management.
     b. Enter the name of the user or group you wish to authorize in the Search field.
     c. Select the user or group from the list that displays.
     d. Click Roles Update Roles .
     e. From Update Roles, select DFCatalogAdmin and click Update.

        Note: If the ReadyFlow is already in the Catalog, then you can give your user just the DFCatalogViewer role.
  2. Give your user or group permission to deploy flow definitions.

     a. From the Management Console, click Environments to display the Environment List page.
     b. Select the environment to which you want your user or group to deploy flow definitions.
     c. Click Actions Manage Access to display the Environment Access page.
     d. Enter the name of your user or group you wish to authorize in the Search field.
     e. Select your user or group and click Update Roles.
     f. Select DFFlowAdmin from the list of roles.
     g. Click Update Roles.
  3. Give your user or group access to the Project where the ReadyFlow will be deployed.

     a. Go to DataFlow Projects .
     b. Select the project where you want to manage access rights and click ⋮ More Manage Access .
  4. Start typing the name of the user or group you want to add and select them from the list.
  5. Select the Resource Roles you want to grant.
  6. Click Update Roles.
  7. Click Synchronize Users.
- You have enabled Inbound Connection Support for your DataFlow during flow deployment.

  For more information, see Create an Inbound Connection Endpoint during flow deployment .

For your ADLS data ingest target

- You have your ADLS container and path into which you want to ingest data.

- You have performed one of the following to configure access to your ADLS folder:
  - You have configured access to the ADLS folders with a RAZ enabled environment.

    It is a best practice to enable RAZ to control access to your object store folders. This allows you to use your CDP credentials to access ADLS folders, increases auditability, and makes object store data ingest workflows portable across cloud providers.

    1. Ensure that Fine-grained access control is enabled for your DataFlow environment.
    2. From the Ranger UI, navigate to the ADLS repository.
    3. Create a policy to govern access to the ADLS container and path used in your ingest workflow. For example: adls-to-adls-avro-ingest

       Tip: The Path field must begin with a forward slash (/).
    4. Add the machine user that you have created for your ingest workflow to ingest the policy you just created.

    For more information, see *Ranger policies for RAZ-enabled Azure environment*.
  - You have configured access to ADLS folders using ID Broker mapping.

    If your environment is not RAZ-enabled, you can configure access to ADLS folders using ID Broker mapping.

    1. Access IDBroker mappings.
       1. To access IDBroker mappings in your environment, click Actions Manage Access .
       2. Choose the IDBroker Mappings tab where you can provide mappings for users or groups and click Edit.
    2. Add your CDP Workload User and the corresponding Azure role that provides write access to your folder in ADLS to the Current Mappings section by clicking the blue + sign.

       Note: You can get the Azure Managed Identity Resource ID from the Azure Portal by navigating to Managed Identities Your Managed Identity Properties Resource ID . The selected Azure MSI role must have a trust policy allowing IDBroker to assume this role.
    3. Click Save and Sync.

For your S3 data ingest target

- You have your source S3 path and bucket.

- Perform one of the following to configure access to S3 buckets:
  - You have configured access to S3 buckets with a RAZ enabled environment.

    It is a best practice to enable RAZ to control access to your object store buckets. This allows you to use your CDP credentials to access S3 buckets, increases auditability, and makes object store data ingest workflows portable across cloud providers.

    1. Ensure that Fine-grained access control is enabled for your DataFlow environment.
    2. From the Ranger UI, navigate to the S3 repository.
    3. Create a policy to govern access to the S3 bucket and path used in your ingest workflow.

       Tip:

       The Path field must begin with a forward slash (/).
    4. Add the machine user that you have created for your ingest workflow to the policy you just created.

    For more information, see *Creating Ranger policy to use in RAZ-enabled AWS environment*.
  - You have configured access to S3 buckets using ID Broker mapping.

    If your environment is not RAZ-enabled, you can configure access to S3 buckets using ID Broker mapping.

    1. Access IDBroker mappings.
       a. To access IDBroker mappings in your environment, click Actions Manage Access .
       b. Choose the IDBroker Mappings tab where you can provide mappings for users or groups and click Edit.
    2. Add your CDP Workload User and the corresponding AWS role that provides write access to your folder in your S3 bucket to the Current Mappings section by clicking the blue + sign.

       Note: You can get the AWS IAM role ARN from the Roles Summary page in AWS and can copy it into the IDBroker role field. The selected AWS IAM role must have a trust policy allowing IDBroker to assume this role.
    3. Click Save and Sync.

Related Concepts

List of required configuration parameters for the Slack to S3/ADLS ReadyFlow

List of required configuration parameters for the Slack to S3/ADLS ReadyFlow

When deploying the Slack to S3/ADLS ReadyFlow, you have to provide the following parameters. Use the information you collected in *Prerequisites*.

Table 1: Slack to S3/ADLS ReadyFlow configuration parameters

Parameter name	Description
Allowed Paths	Specify the allowed HTTP paths configured in your Slack App. The default value (/events) allows Slack events.
CDP Workload User	Specify the CDP machine user or workload username that you want to use to authenticate to the object stores. Ensure this user has the appropriate access rights to the object store locations in Ranger or IDBroker.
CDP Workload User Password	Specify the password of the CDP machine user or workload user you are using to authenticate against the object stores (via IDBroker).
CSV Delimiter	If your desired output data is CSV, specify the delimiter here.
Data Output Format	Specify the desired format for your output data. You can use CSV, JSON or AVRO with this ReadyFlow.

Parameter name	Description
Destination S3 or ADLS Path	Specify the name of the destination S3 or ADLS path you want to write to. Make sure that the path starts with "/".
Destination S3 or ADLS Storage Location	Specify the name of the destination S3 bucket or ADLS container you want to write to. • For S3, enter a value in the form: s3a://*[***Destination S3 Bucket***]* • For ADLS, enter a value in the form: abfs://*[***Destination ADLS File System***]*@*[***Destination ADLS Storage Account***]*.dfs.core.windows.net
Listening Port	Specify the port to listen on for incoming connections.

Related Concepts

Prerequisites

Related Information

Deploying a ReadyFlow