

Azure Event Hub to ADLS

Date published: 2021-04-06

Date modified: 2024-06-03

CLOUdera

Legal Notice

© Cloudera Inc. 2024. All rights reserved.

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: Azure Event Hub to ADLS.....	4
Prerequisites.....	4
List of required configuration parameters for the Azure Event Hub to ADLS ReadyFlow.....	8

ReadyFlow Overview: Azure Event Hub to ADLS

You can use the Azure Event Hub to ADLS ReadyFlow to move JSON, CSV or Avro files from an Azure Event Hub namespace, optionally parsing the schema using Cloudera Schema Registry or direct schema input. The flow then filters records based on a user-provided SQL query and writes them to a target Azure Data Lake Storage (ADLS) location in the specified output data format.

This ReadyFlow consumes JSON, CSV or Avro data from a source Azure Event Hub Namespace and merges the events into JSON, CSV or Avro files before writing the data to ADLS. 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. You can specify the Event Hub Namespace you want to read from as well as the target ADLS data container and path. Failed ADLS write operations are retried automatically to handle transient issues. Define a KPI on the failure_WriteToADLS connection to monitor failed write operations.



Note: This ReadyFlow leverages Cloudera Public Cloud'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.

Azure Event Hub to ADLS ReadyFlow details	
Source	Azure Event Hub Namespace
Source Format	JSON, CSV, Avro
Destination	ADLS
Destination Format	JSON, CSV, Avro

Moving data to object stores

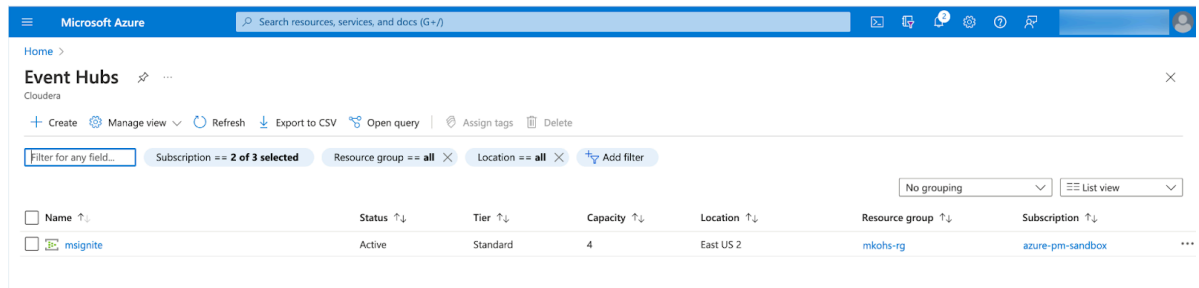
Cloud environments offer numerous deployment options and services. There are many ways to store data in the cloud, but the easiest option is to use object stores. Object stores are extremely robust and cost-effective storage solutions with multiple levels of durability and availability. You can include them in your data pipeline, both as an intermediate step and as an end state. Object stores are accessible to many tools and connecting systems, and you have a variety of options to control access.

Prerequisites

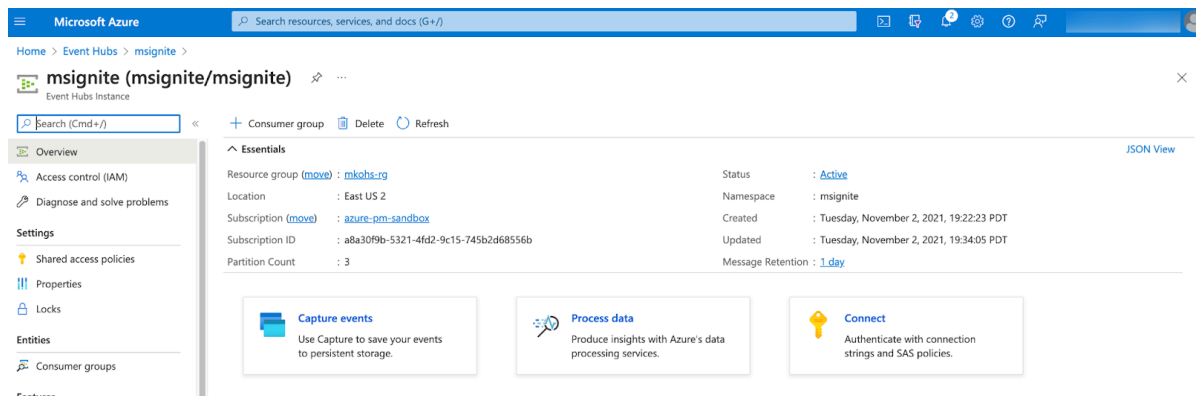
Learn how to collect the information you need to deploy the Azure Event Hub to ADLS ReadyFlow, and meet other prerequisites.

For your data ingest source

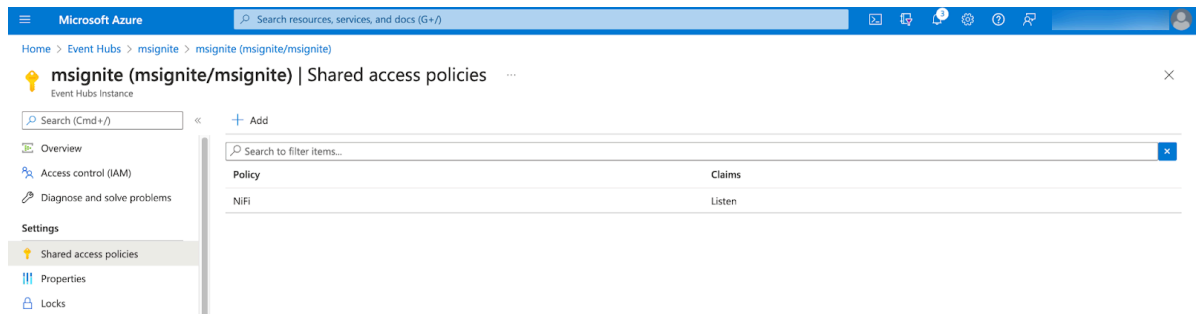
- You have obtained the Event Hub Namespace and Instance name
- Navigate to **Event Hubs** and select your Event Hub instance (msignite, in this example):



- Copy the namespace and the Event Hub Instance name (msignite/msignite in this example).

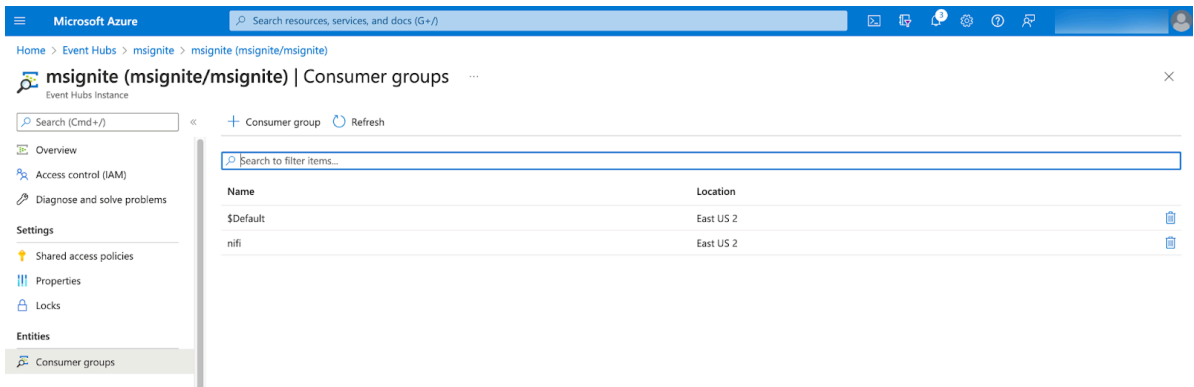


- You have created an Event Hub Shared Access Policy and obtained its name and Primary Access Key
- Click **Shared access policies**.
- Set up a shared access policy that allows "Listen". Copy the Policy name.



- Select the policy and copy the primary key.

- You have created a consumer group.
- Click Consumer groups.
- Use the \$Default consumer group or create your own group for the ReadyFlow:



For Cloudera DataFlow

- You have enabled Cloudera DataFlow for an environment.

For information on how to enable Cloudera DataFlow for an environment, see [Enabling Cloudera DataFlow for an Environment](#).

- You have created a Machine User to use as the Cloudera Workload User.
- You have given the Cloudera Workload User the EnvironmentUser role.
 1. From the Management Console, go to the environment for which Cloudera 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 Cloudera 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 Cloudera Public Cloud environment that you enabled for Cloudera DataFlow.

For information on how to synchronize your user to FreeIPA, see [Performing User Sync](#).

- You have granted your Cloudera 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.

For your 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 Cloudera Public Cloud 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 Cloudera 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.
 - 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 Cloudera 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.

Related Concepts

[List of required configuration parameters for the Azure Event Hub to ADLS ReadyFlow](#)

List of required configuration parameters for the Azure Event Hub to ADLS ReadyFlow

When deploying the Azure Event Hub to ADLS ReadyFlow, you have to provide the following parameters. Use the information you collected in *Prerequisites*.

Azure Event Hub to ADLS ReadyFlow configuration parameters	
Parameter Name	Description
ADLS File System	Specify the name of the ADLS data container you want to write to. The full path will be constructed from: abfs://{ADLS File System}@#{ADLS Storage Account}.dfs.core.windows.net/#{ADLS Path}/\${Kafka.topic}
ADLS Path	Specify the path within the ADLS data container where you want to write to without any leading characters. The full path will be constructed from: abfs://{ADLS File System}@#{ADLS Storage Account}.dfs.core.windows.net/#{ADLS Path}/\${Kafka.topic}

Azure Event Hub to ADLS ReadyFlow configuration parameters	
Parameter Name	Description
ADLS Storage Account	Specify the storage account name you want to write to. The full ADLS data container path will be constructed from: abfs://{ADLS File System}@#{ADLS Storage Account}.dfs.core.windows.net/#{ADLS Path}/{Kafka.topic}
CDP Workload User	Specify the Cloudera machine user or workload username that you want to use to authenticate to Kafka and the object store. Ensure this user has the appropriate access rights in Ranger for the Kafka topic and Ranger or IDBroker for object store access.
CDP Workload User Password	Specify the password of the Cloudera machine user or workload user you are using to authenticate against Kafka and the object store.
CSV Delimiter	If your source data is CSV, specify the delimiter here.
Data Output Format	Specify the desired format for your output data. You can select from <ul style="list-style-type: none"> • CSV • JSON • AVRO with this ReadyFlow.
Event Hub Access Policy Name	Specify the Access Policy Name that this flow should use. The full path for the event hub endpoint will be constructed from sb://{Event Hub Namespace}.#{Event Hub Service Bus Endpoint};SharedAccessKeyName=#{Event Hub Access Policy Name};SharedAccessKey=#{Event Hub Access Primary Key}
Event Hub Access Primary Key	Specify the Primary Key that allows clients to use the Access Policy that you provided earlier. The full path for the event hub endpoint will be constructed from sb://{Event Hub Namespace}.#{Event Hub Service Bus Endpoint};SharedAccessKeyName=#{Event Hub Access Policy Name};SharedAccessKey=#{Event Hub Access Primary Key}
Event Hub Consumer Group	Specify the Event Hub Consumer Group you want to use with this flow. Any consumer group other than \$Default needs to be created in Event Hub first.
Event Hub Instance Name	Specify the Event Hub Instance Name inside the Event Hub Namespace you want to use.
Event Hub Namespace	Specify the Event Hub Namespace which contains the Event Hub instance you want to use. The full path for the event hub endpoint will be constructed from sb://{Event Hub Namespace}.#{Event Hub Service Bus Endpoint};SharedAccessKeyName=#{Event Hub Access Policy Name};SharedAccessKey=#{Event Hub Access Primary Key}
Event Hub Partitions Count	Specify the number of partitions that the Event Hub has. Only this number of partitions will be used, so it is important to ensure that if the number of partitions changes that this value be updated. Otherwise, some messages may not be consumed.
Event Hub Service Bus Endpoint	Specify the Event Hub Service Bus Endpoint. The default value is .servicebus.windows.net The full path for the event hub endpoint will be constructed from sb://{Event Hub Namespace}.#{Event Hub Service Bus Endpoint};SharedAccessKeyName=#{Event Hub Access Policy Name};SharedAccessKey=#{Event Hub Access Primary Key}
Filter Rule	Specify the filter rule expressed in SQL to filter streaming events for the destination object store. Records matching the filter will be written to the destination object store. The default value forwards all records.
Schema Text	Specify the Avro-formatted schema to be used for the source event hub data.

Related Concepts

[Prerequisites](#)

[Related Information](#)

[Deploying a ReadyFlow](#)