Cloudera DataFlow for Data Hub 7.2.11

Creating your first Streaming Analytics cluster in CDP Public Cloud

Date published: 2019-12-16
Date modified: 2021-09-09

https://docs.cloudera.com/

Legal Notice

Contents

Creating your first Streaming Analytics cluster

You can create a managed and secured Streaming Analytics cluster in a CDP Public Cloud environment by using one of the prescriptive streaming analytics cluster definitions and following the cluster creation steps in Data Hub.

Meet the prerequisites

Before you start creating your first Streaming Analytics Data Hub cluster, you need to ensure that you have set up the environment properly and have all the necessary accesses to use CDP Public Cloud.

- You have CDP login credentials.
- You have an available CDP environment.
- You have a running Data Lake.
- You have a CDP username and the predefined resource role of this user is EnvironmentAdmin.
- Your CDP user is synchronized to the CDP Public Cloud environment.

Important: Ensure that the Runtime version of the Data Lake cluster matches the Runtime version of the Data Hub cluster that you are about to create. If these versions do not match, you may encounter warnings and/or errors.

Related Information

Getting started as a user
AWS environments
Azure environments
GCP environments
Data lakes

Create your cluster

When creating your Streaming Analytics cluster, you must choose from the Light and Heavy duty options. You also need to pay attention to the cloud storage settings where Flink saves the checkpoints and savepoints.

About this task

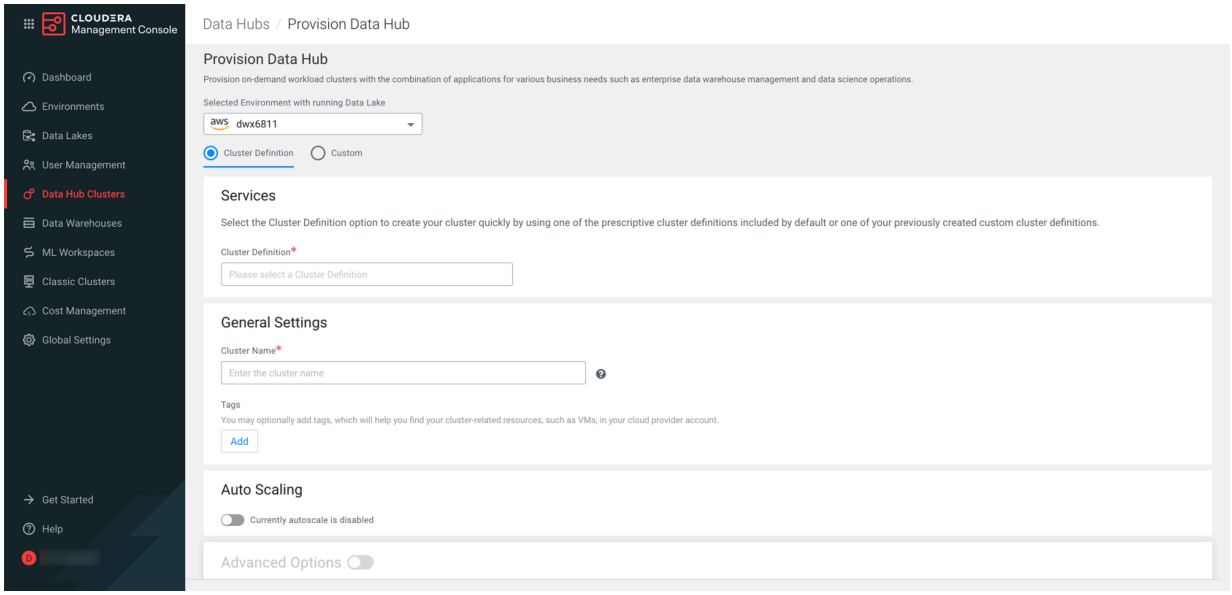
After you have met the prerequisites, you are ready to create your Streaming Analytics cluster using a default cluster definition.

Procedure

1. Log into the CDP web interface.
2. Navigate to Management Console > Environments , and select the environment where you want to create a cluster.

3. Click Create Data Hub.

   The following page is displayed:



4. Select Cluster Definition.
5. Select the Streaming Analytics cluster definition from the Cluster Definition drop-down list depending on your operational objectives.

   You have the following options:

   - 7.2.11 - Streaming Analytics Light Duty for AWS
   - 7.2.11 - Streaming Analytics Light Duty for Azure
   - 7.2.11 - Streaming Analytics Light Duty for GCP
   - 7.2.11 - Streaming Analytics Heavy Duty for AWS
   - 7.2.11 - Streaming Analytics Heavy Duty for Azure
   - 7.2.11 - Streaming Analytics Heavy Duty for GCP

   For more information on templates, see Streaming Analytics deployment scenarios.

   The list of services is automatically shown below the selected cluster definition name.
6. Provide a cluster name and add tags you might need.

   You can define tags that will be applied to your cluster- related resources on your cloud provider account. For more information about tags, see Tags.
7. Optionally, use the Configure Advanced Options section to customize the infrastructure settings.

   Note: Ensure that the right cloud storage path is given in Advanced Options > Cloud Storage . Cloudera recommends saving the checkpoints and savepoints to the S3 cloud storage to make the saved files available throughout all cluster deployments. You can also use HDFS, however Cloudera only recommends this solution for temporary storage of checkpoints and savepoints.
8. Click Provision Cluster.

Results

You are redirected to the Data Hub cluster dashboard, and a new tile representing your cluster appears at the top of the page.

Related Information

Create a cluster from a definition on AWS

Create a cluster from a definition on Azure

Create a cluster from a definition on GCP

Tags
Streaming Analytics deployment scenarios

Give users access to your cluster

As an EnvironmentAdmin, you need provide access to users to your environment and to the Streaming Analytics cluster by assigning user roles, adding users to Ranger policies, and creating IDBroker mappings.

About this task

The cluster you have created using the Streaming Analytics cluster definition is kerberized and secured with SSL. Users can access cluster UIs and endpoints through a secure gateway powered by Apache Knox. Before you can use Flink and SQL Stream Builder, you must provide users access to the Streaming Analytics cluster components.

Procedure

1. Assign the EnvironmentUser role to the users to grant access to the CDP environment and the Streaming Analytics cluster.
2. Add the user to the appropriate predefined Ranger policies.
3. Create IDBroker mapping.

   You must create IDBroker mapping for a user or group to have access to the S3 cloud storage. As a part of Knox, the IDBroker allows a user to exchange cluster authentication for temporary cloud credentials.

   The following roles are created when registering the CDP environment:

   - idbroker-role: granting permissions to IDBroker istances associated wit the CDP environemtn
   - datalake-admin-role: granting access to CDP cloud rescources
   - logs-role: granting access to the logs storage location

   For using Streaming Analytics in CDP Public Cloud, you must map the users using Flink to the the datalake-adm in-role as it grants access to the cloud resources required to run the Flink service. You can configure the IDBroker mappings for Flink users with the following steps:

   a. Navigate to Management Console > Environments , and select the environment where you have created your cluster.
   b. Click on Actions > Manage Access .
   c. Select IDBroker Mappings tab.
   d. Click Edit in the Current Mappings pane.
   e. Make sure that the users who run Flink jobs are associated with the ARN of the datalake-admin-role.

      If you need to add more users or groups:

      1. Add a new row using the plus icon.
      2. Search and select the user or group.
      3. Click Save and Sync.

Related Information

IDBroker
Set Ranger policies for Flink
Set Ranger policies for SQL Stream Builder

Next steps

After you have created your Streaming Analytics data hub cluster, you are ready to design your first streaming application in CDP Public Cloud. You can further gather information about Flink, streaming application and SQL queries through other documentation to get familiar with the analytical use cases.

Related Information

Creating a Flink Streaming Application

DataStream Connectors

Configuring Flink applications for production

Analyzing data using Apache Flink

Flink Application Tutorials

Registering Data Providers in SSB

Using Tables in SQL Stream jobs

Querying data using SQL Stream Builder