Cloudera DataFlow for Data Hub 7.2.18

Setting up your Streaming Analytics cluster

Date published: 2019-12-16
Date modified: 2024-04-03

https://docs.cloudera.com/

Legal Notice

Contents

Before creating your cluster

Before you start creating your 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

Assigning resource roles

As an administrator, you need to give permissions to users or groups to be able to access and perform tasks in your Data Hub environment.

Procedure

1. Navigate to Management Console > Environments and select your environment.
2. Click Actions > Manage Access .
3. Search for a user or group that needs access to the environment.
4. Select EnvironmentUser role from the list of Resource Roles.
5. Click Update Roles.
   The Resource Role for the selected user or group will be updated.
6. Navigate to Management Console > Environments , and select the environment where you want to create a cluster.
7. Click Actions > Synchronize Users to FreeIPA .
8. Click Synchronize Users.

Note: There might be cases where the status of the environment is synchronized with warnings and has failed status. This does not indicate that the synchronization has failed.

Creating IDBroker mapping

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

About this task

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 instances associated with the CDP environment
- datalake-admin-role: granting access to CDP cloud resources
- logs-role: granting access to the logs storage location

For using Streaming Analytics in CDP Public Cloud, you must make sure that the users who run Flink jobs are associated with the ARN of the datalake-admin-role as it grants access to the cloud resources required to run the Flink service.

Procedure

1. Navigate to Management Console > Environments and select your environment.
2. Click Actions > Manage Access .
3. Click on the IDBroker Mappings tab.
4. Click Edit to add a new user or group and assign roles to have writing access for the cloud storage.
5. Search for the user or group you need to map.
6. Go to the IAM Summary page where you can find information about your cloud storage account.
7. Copy the Role ARN.
8. Go back to the IDBroker Mapping interface on the Cloudera Management Console page.
9. Paste the Role ARN to your selected user or group.
10. Click Save and Sync.

Setting workload password

As a user, you need to set a workload password for your EnvironmentUser account to be able to access the SQL Stream Builder nodes through SSH connection.

Procedure

1. Navigate to Management Console > Environments and select your environment.
2. Click Actions > Manage Access .
3. Click Workload Password.
4. Give a chosen workload password for your user.
5. Confirm the given password by typing it again.
6. Click Set Workload Password.

Creating your cluster

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

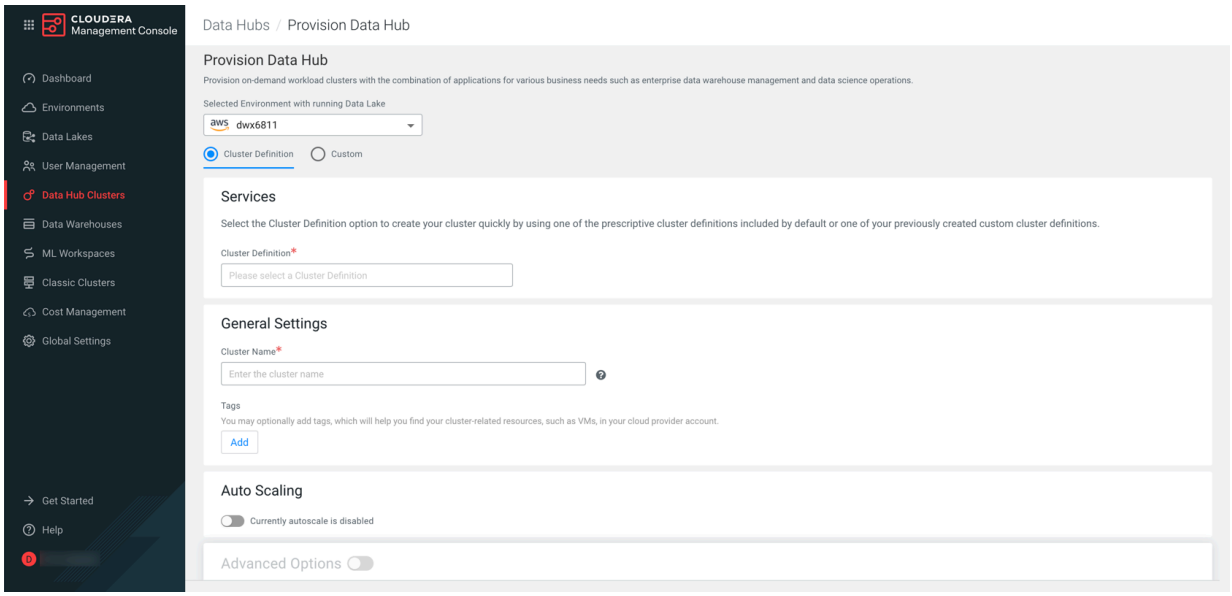
About this task

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.

   The following template options are available:

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

   For more information on templates, see *Streaming Analytics Data Hub cluster definitions* and *Streaming Analytics cluster layout*.

   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 Data Hub cluster definitons
Streaming Analytics cluster layout

After creating your cluster

As an EnvironmentAdmin, you need to 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.

Related Information

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

Retrieving keytab file

As a user, you need to retrieve the keytab file of your profile and upload it to the Streaming SQL Console to be able to run SQL jobs.

Procedure

1. Navigate to Management Console > Environments , and select the environment where you have created your cluster.
2. Click on your profile name.
3. Click Profile.
4. Click Actions > Get Keytab.
5. Choose the environment where your Data Hub cluster is running.
6. Click Download.
7. Save the keytab file in a chosen location.

Uploading or unlocking your keytab

When accessing the Streaming SQL Console for the first time in Data Hub, you must upload and unlock the keytab file corresponding with your profile before you can use SQL Stream Builder (SSB).

Procedure

1. Navigate to the Streaming SQL Console.
   a) Navigate to Management Console > Environments , and select the environment where you have created your cluster.
   b) Select the Streaming Analytics cluster from the list of Data Hub clusters.
   c) Select Streaming SQL Console from the list of services.

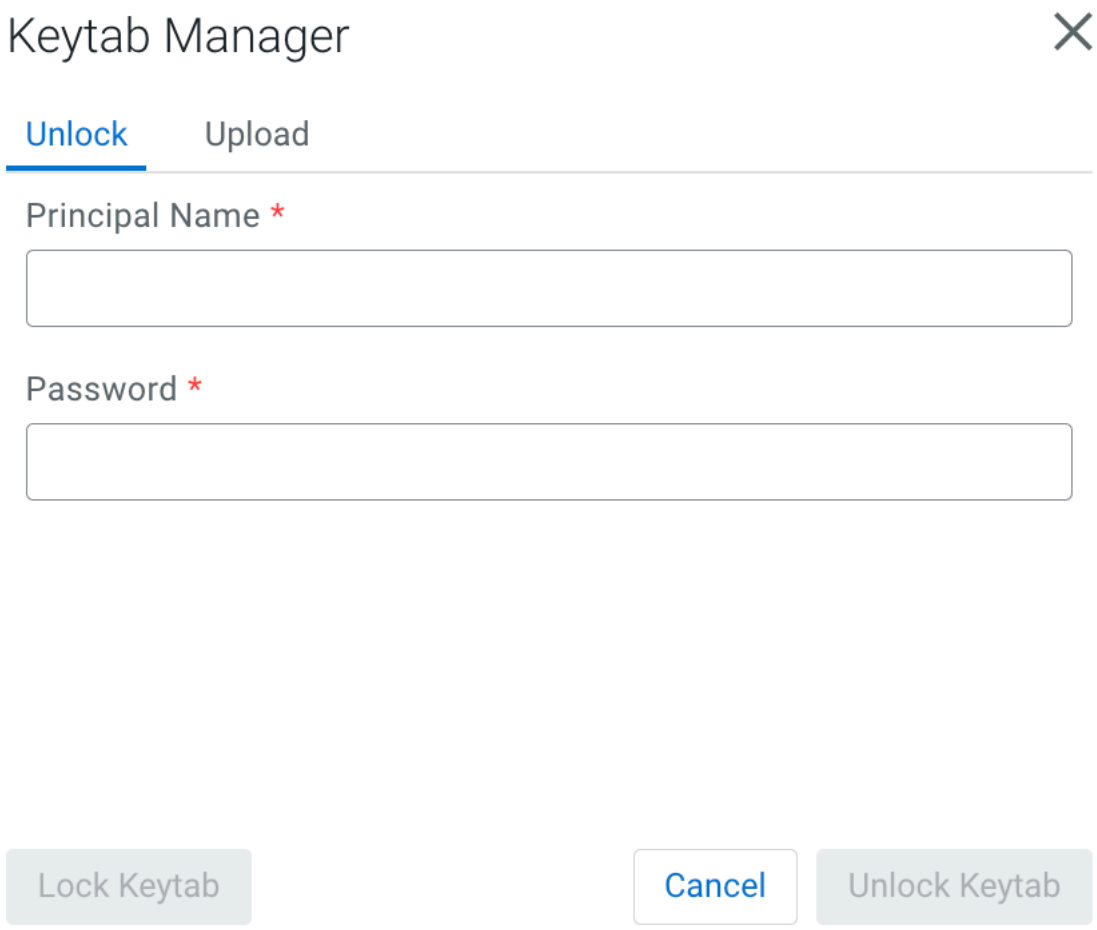
      The **Streaming SQL Console** opens in a new window.
2. Click your username on the sidebar of the Streaming SQL Console.

3. Click Manage keytab.

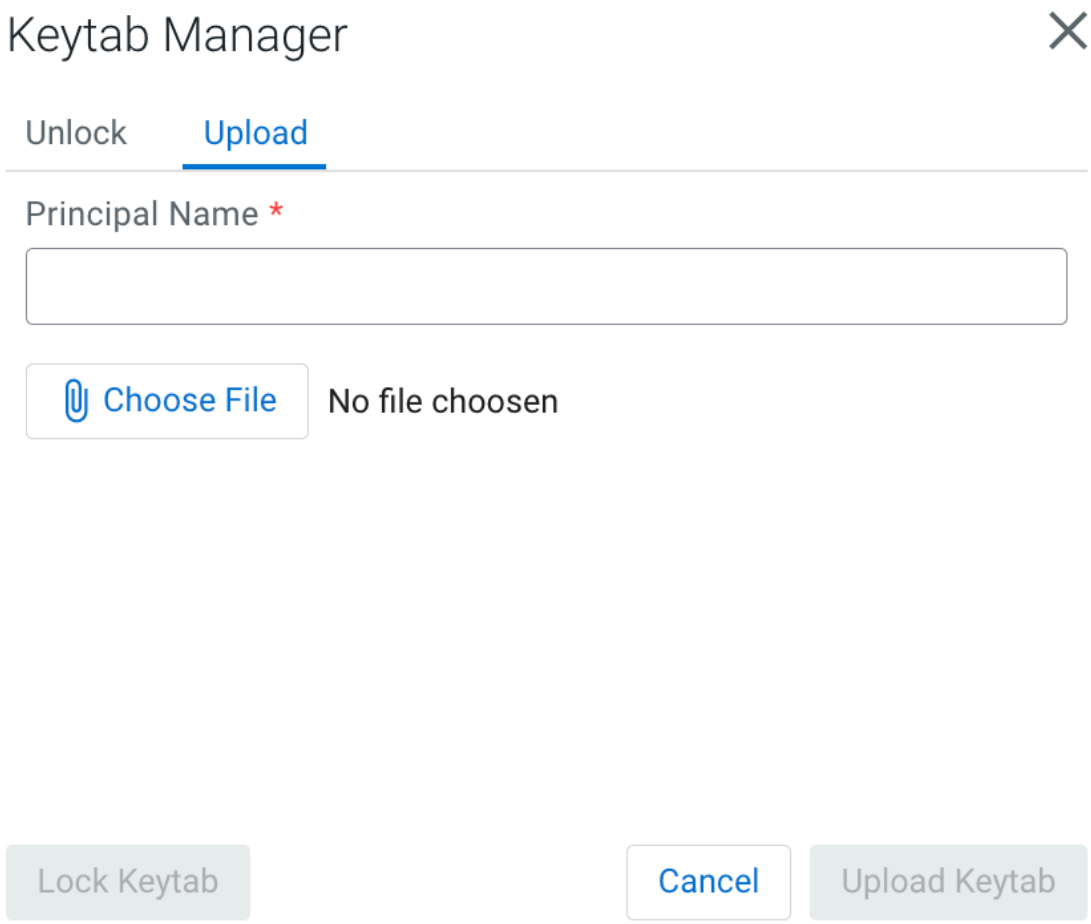
   The **Keytab Manager** window appears.

   You can either unlock the keytab already existing on the cluster, or you can directly upload your keytab file in the SQL Stream Builder.

   a) Unlock your keytab by providing the Principal Name and Password, and clicking Unlock Keytab. The Principal Name and Password should be the same as the workload username and password set for the Streaming Analytics cluster.



   b) Upload your keytab by clicking on the Upload tab, uploading the keytab file directly to the Console, and clicking Unlock Keytab.



In case there is an error when unlocking your keytab, you can get more information about the issue with the following steps:

a. Retrieve your keytab file.

1. Click on your profile name in the **Management Console**.
2. Click Profile.
3. Click Actions > Get Keytab.
4. Choose the environment where your Data Hub cluster is running.
5. Click Download.
6. Save the keytab file in a chosen location.

b. Manually upload your keytab to the Streaming Analytics cluster:

```
scp <location>/<your_keytab_file> <workload_username>@<manager_node_FQDN
>:.
                    Password:<your_workload_password>
```

c. Access the manager node of your Streaming Analytics cluster:

```
ssh <workload_username>@<manager_node_FQDN>
```

```
                                Password: <workload_password>
```

d. Use kinit command to authenticate your user:

```
kinit -kt <keytab_filename>.keytab <workload_username>
```

e. Use the flink-yarn-session command to see if the authentication works properly:

```
flink-yarn-session -d \
-D security.kerberos.login.keytab=<keytab_filename>.keytab \
-D security.kerberos.login.principal=<workload_username>
```

In case the command fails, you can review the log file for further information about the issue.

Configuring Ranger policies for Flink and SSB

You must add your workload username and the SQL Stream Builder (SSB) service user to the Ranger policies that are needed for Kafka, Schema Registry, Hive and Kudu to provide access to topics, schemas and tables used by the components and to be able to execute Flink jobs.

About this task

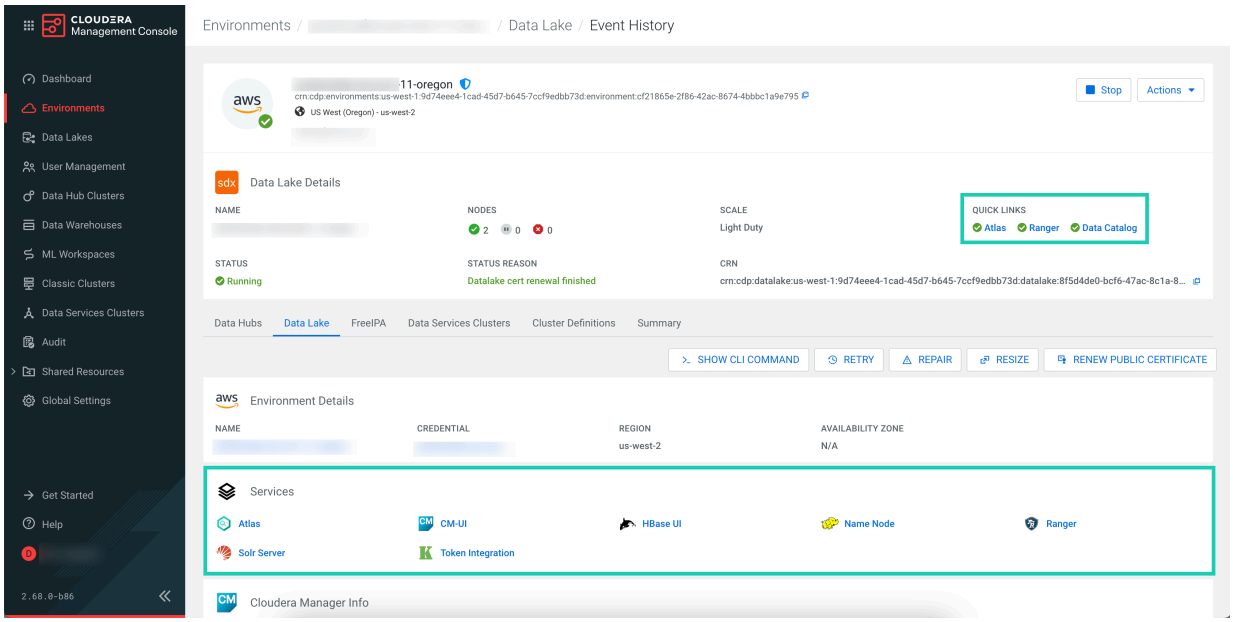
You need to provide access to users and the SSB service by configuring Ranger policies for the Kafka data source and the Schema Registry, Kudu and Hive catalog services. To be able to use Flink, you need to add the workload user or users to the required policies. For SSB, the ssb service user needs to be added to the same policies.

When adding more workload users, instead of adding them one by one, you can create user groups in Ranger, for example a flink_users group. This way you can assign every user who will use the Streaming Analytics cluster into a group, and add only that one group to the Ranger policies.

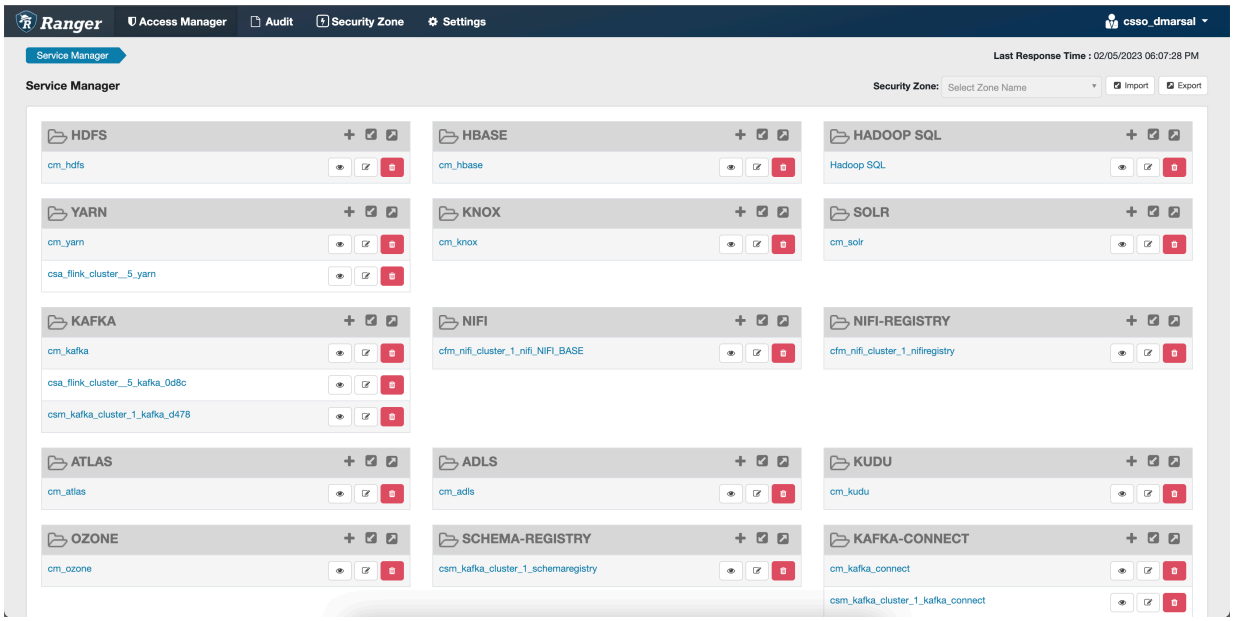
Procedure

1. Navigate to Management Console > Environments , and select the environment where you have created your cluster.

2. Click Ranger from the **Quick Links** or select Data Lake Ranger from **Services**.



You are redirected to the Ranger Admin Web user interface (UI) where you can add the workload user and SSB service user to the required policies to grant access for Flink and SSB.



Configuring Kafka policies

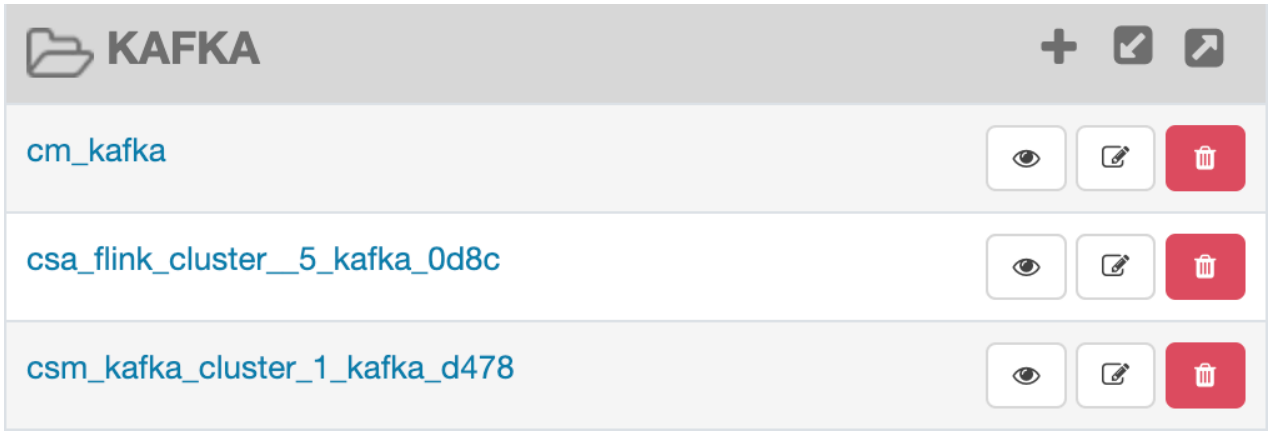
After accessing the Ranger Admin Web UI, the workload username or user groups, and the SSB service user needs to be added to the Kafka policies to be able to use Flink and SSB.

About this task

The following resource based policies need to be configured for the Kafka service:

- all - topic: Provides access to all topics for users
- all - consumergroup: Provides access to all consumer groups for users
- all - cluster: Provides access to all clusters to users
- all - transactionalid: Provides transactionalid access to users
- all - delegationtoken: Provides delegationtoken access to users

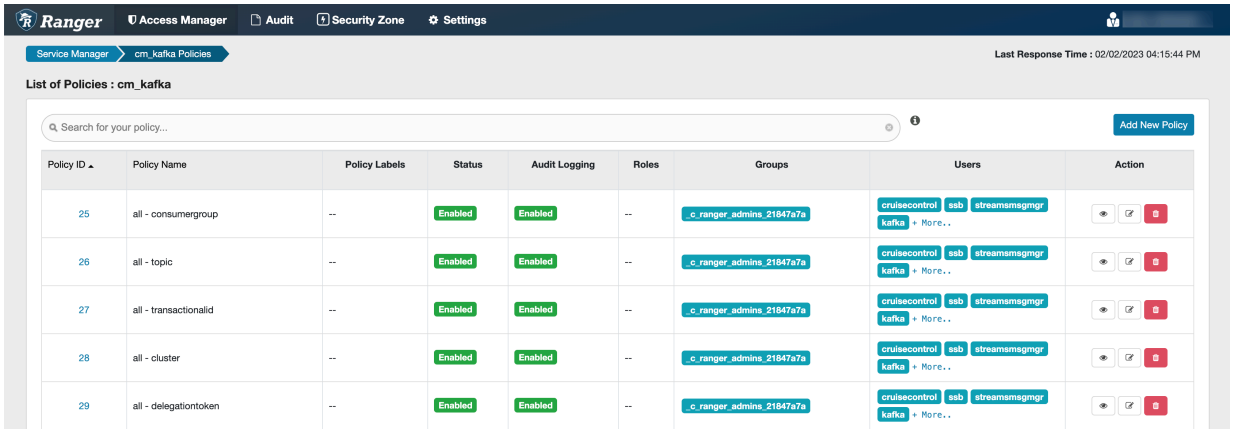
You need to ensure that the required workload username or user group, and the ssb service user is added to the policies of the Kafka service and to the policies of the created Streams Messaging and Streaming Alaytics clusters.



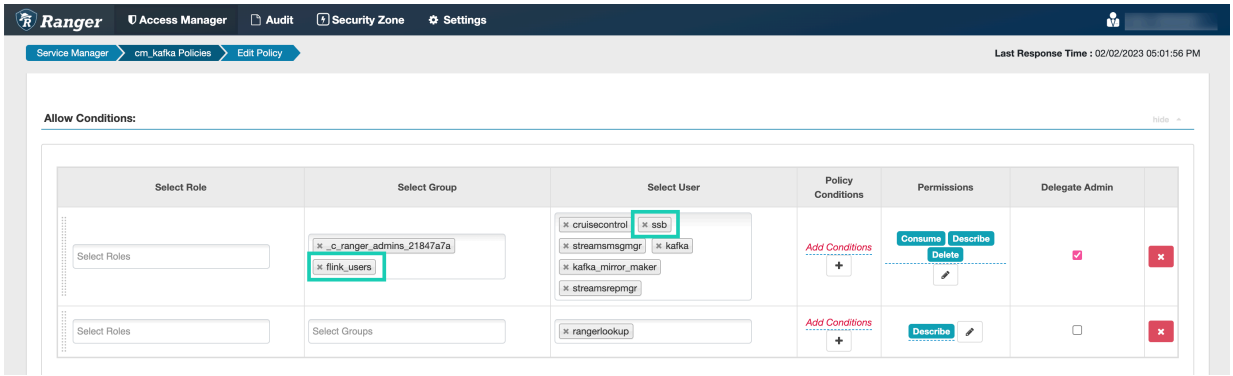
Procedure

1. Click cm_kafka under **Kafka** service on the **Service Manager** page.

   You are redirected to the **cm_kafka Policies** page.



2. Click on the edit button of the *all-consumergroup* policy.
3. Add the user group to the Select Group field under the **Allow Conditions** settings.
   - Alternatively, you can also add the workload usernames to the Select User field under the **Allow Conditions** setting.
4. Add the ssb service user to the Select User field under the **Allow Conditions** setting, if it is not configured by default.



5. Click Save.

   You are redirected to the list of Kafka policies page.

6. Click on + More… to check if the needed workload user is listed under the **Users** for the edited policy.

   Repeate same steps to the remaining Kafka policies as well based on the required authorization level:

   - all - topic
   - all - transactionalid
   - all - cluster

Configuring Schema Registry policies

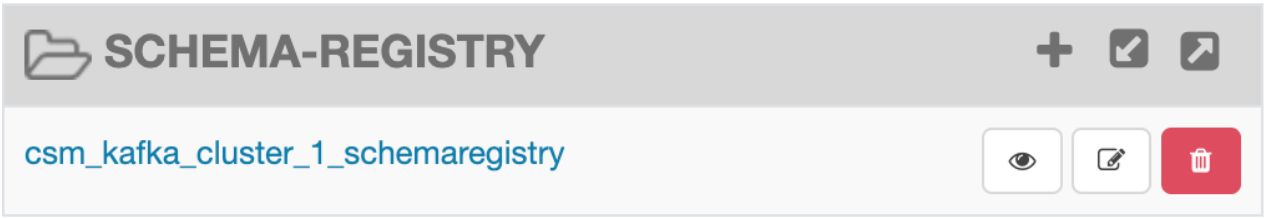
After accessing the Ranger Admin Web UI, the workload username or user groups, and the SSB service user needs to be added to the Schema Registry policies to be able to use Flink and SSB.

About this task

The following resource based policy need to be configured for the Schema Registry service:

- all - export-import: Provides import and export access for users.
- all - serde: Provides access to store metadata for the format of how the data should be read and written.
- all - schema-group, schema-metadata: Provides access to the schema groups, schema metadata, and schema branch.
- all - schema-group, schema-metadata, schema-branch: Provides access to the schema groups, schema metadata, and schema branch.
- all - schema-group, schema-metadata, schema-branch, schema-version: Provides access to schema groups, schema metadata, schema branch, and schema version.
- all - registry-service: Provides access to the schema registry service, the user can access all Schema Registry entities.

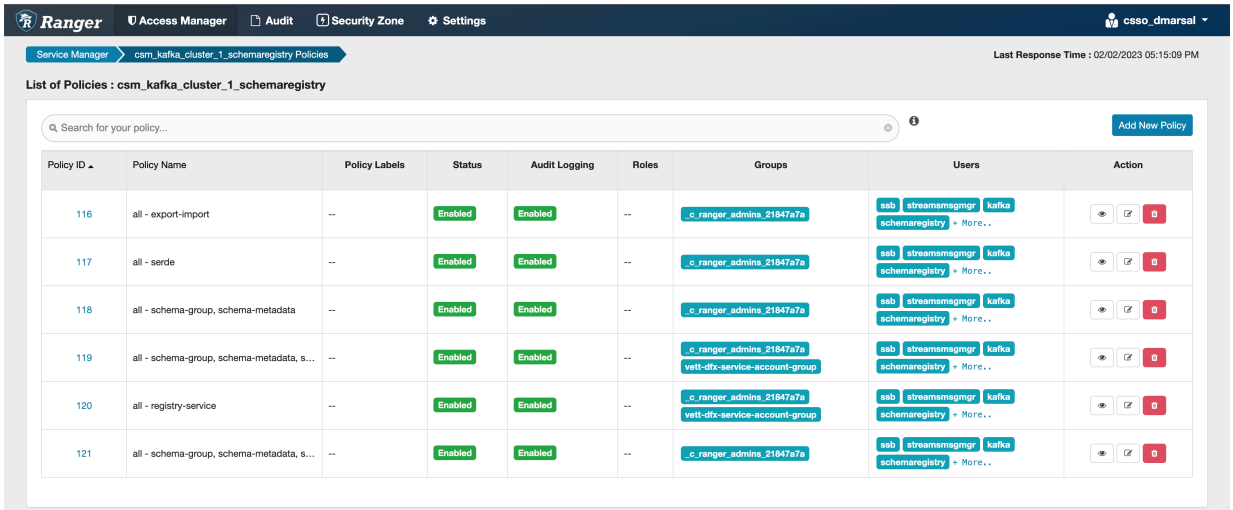
You need to ensure that the required workload username or user group, and the ssb service user is added to the policies of the created Streams Messaging clusters.



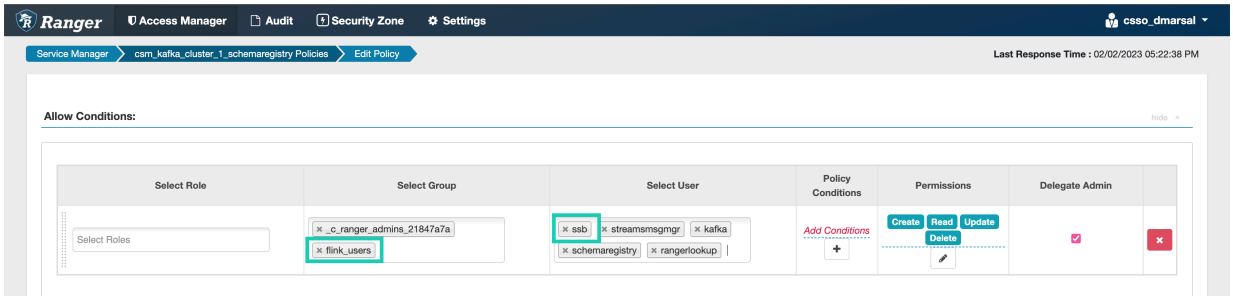
Procedure

1. Select your Streams Messaging cluster under the **Schema Registry** service on the **Service Manager** page.

   You are redirected to the list of Schema Registry policies page.



2. Click on the edit button of the *all-schema-group, schema-metadata, schema-branch, schema-version* policy.
3. Add the user group to the Select Group field under the **Allow Conditions** settings.
   - Alternatively, you can also add the workload usernames to the Select User field under the **Allow Conditions** setting.
4. Add the ssb service user to the Select User field under the **Allow Conditions** setting, if it is not configured by default.



5. Click Save.

   You are redirected to the list of **Schema Registry policies** page.
6. Click on + More… to check if the needed workload user is listed under the **Users** for the edited policy.

   Repeate same steps to the remaining Schema Registry policies as well based on the required authorization level:
   - all - export-import
   - all - serde
   - all - schema-group, schema-metadata
   - all - schema-group, schema-metadata, schema-branch, schema-version
   - all - registry-service

Configuring Hive policies

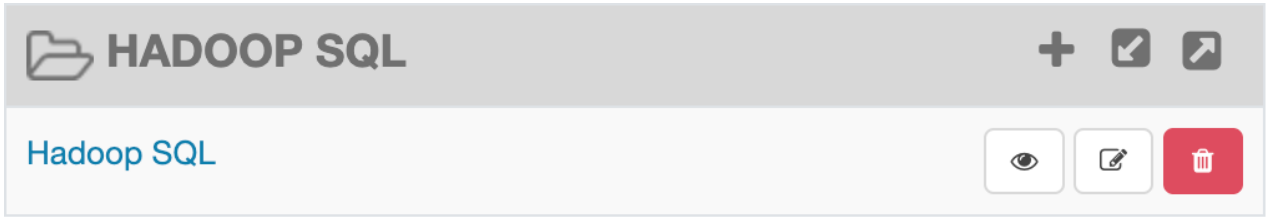
After accessing the Ranger Admin Web UI, the workload username or user groups, and the SSB service user needs to be added to the Hadoop SQL policies to be able to use Flink and SSB with Hive.

About this task

The following resource based policy need to be configured for the Hive service:

- all - global: Provides global access to users.
- all - database, table, column: Provides access to all databases, tables, and columns for users.
- all - database, table: Provides access to all databases and tables for users.
- all - database: Provides access to all databases for users.
- all - hiveservice: Provides hiveservice access to users.
- all - database, udf: Provides database and udf access to users.
- all - url: Provides url access

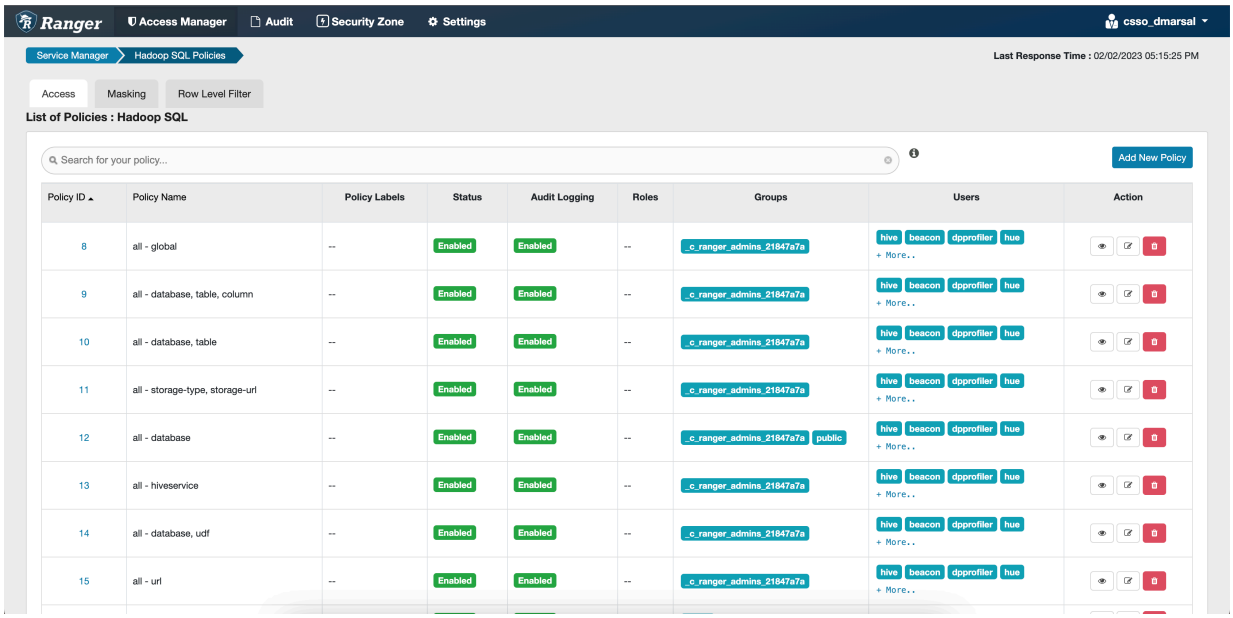
You need to ensure that the required workload username or user group, and the ssb service user is added to the policies of the Hive service and the created Operational Database clusters.



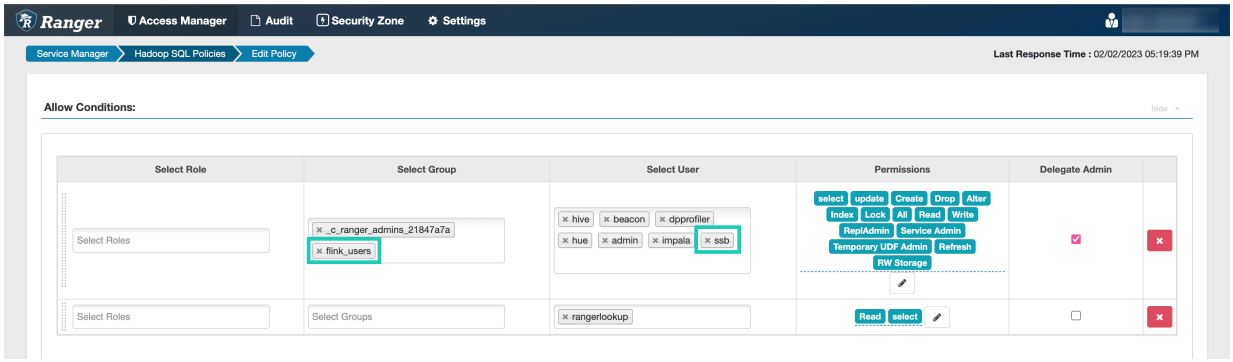
Procedure

1. Click Hadoop SQL under **Hadoop SQL** service on the **Service Manager** page.

   You are redirected to the list of **Hadoop SQL** policies page.



2. Click on the edit button of the *all-global* policy.
3. Add the user group to the Select Group field under the **Allow Conditions** settings.
   - Alternatively, you can also add the workload usernames to the Select User field under the **Allow Conditions** setting.
4. Add the ssb service user to the Select User field under the **Allow Conditions** setting, if it is not configured by default.



5. Click Save.

   You are redirected to the list of **Hadoop SQL policies** page.
6. Click on + More… to check if the needed workload user and the ssb are listed under the **Users** for the edited policy.

   Repeate same steps to the remaining Hadoop SQL policies as well based on the required authorization level:

   - all-database, table, column
   - all-database, table
   - all-database
   - all-database, udf
   - all-hiveservice
   - all-url

Configuring Kudu policies

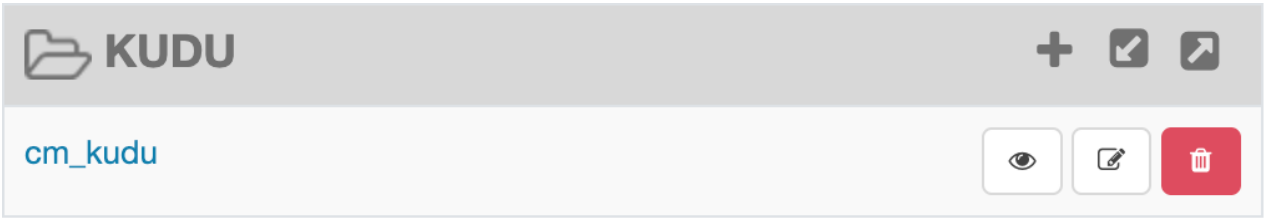
After accessing the Ranger Admin Web UI, the workload username or user groups, and the SSB service user needs to be added to the Kudu policies to be able to use Flink and SSB.

About this task

The following resource based policy need to be configured for the Kudu service:

- all - database, table: Provides access to all databases and tables for users.
- all - database, table, column: Provides access to all databases, tables, and columns for users.
- all - database: Provides access to all databases for users.

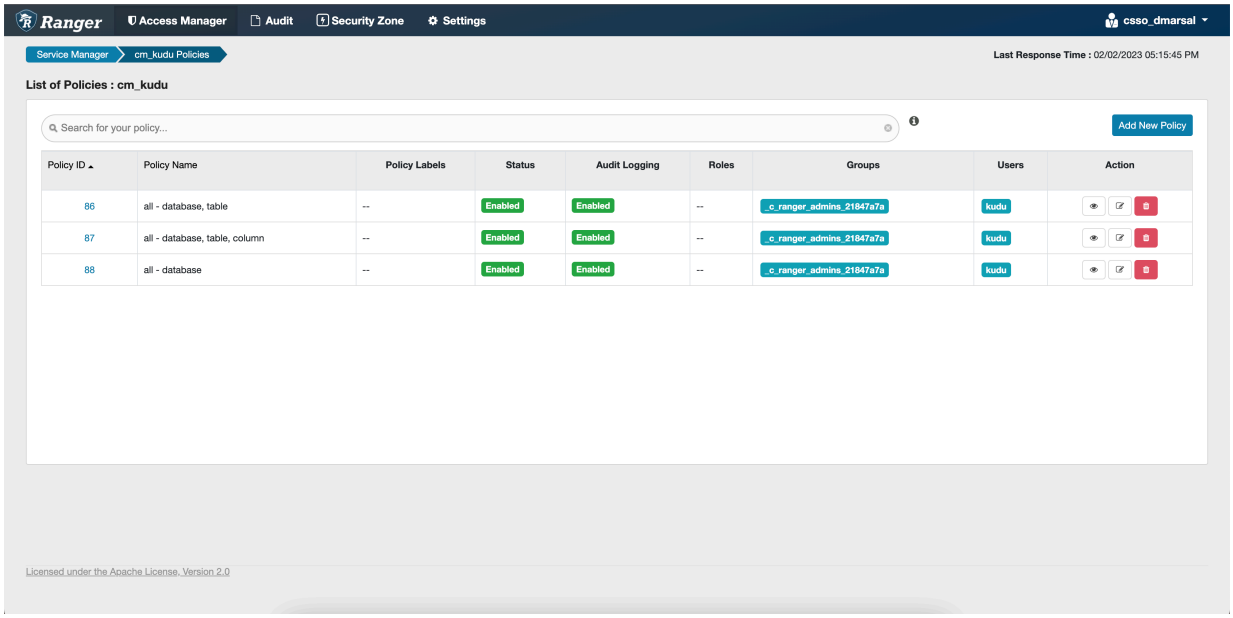
You need to ensure that the required workload username or user group, and the ssb service user is added to the policies of the Kudu service and the created Real-Time Data Mart clusters.



Procedure

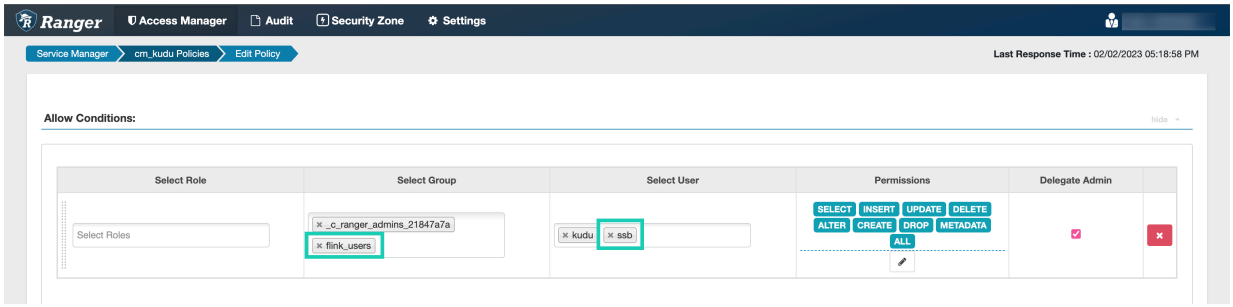
1. Click cm_kudu under **Kudu** service on the **Service Manager** page.

   You are redirected to the **cm_kudu Policies** page.



2. Click on the edit button of the *all-database* policy.
3. Add the user group to the Select Group field under the **Allow Conditions** settings.
   - Alternatively, you can also add the workload usernames to the Select User field under the **Allow Conditions** setting.

4. Add the ssb service user to the Select User field under the **Allow Conditions** setting, if it is not configured by default.



5. Click Save.

   You are redirected to the list of **Kudu policies** page.

6. Click on + More… to check if the needed workload user is listed under the **Users** for the edited policy.

   Repeate same steps to the remaining Kudu policies as well based on the required authorization level:

   - all - database, table
   - all - database, table, column

   Note: From CDP Public Cloud 7.2.16, Kudu environments have their own Ranger policy. Ensure that the permissions are added for every relevant Kudu cluster in the environment.