Cloudera Runtime 7.2.6

Use cases for Streams Replication Manager in CDP Public Cloud

Date published: 2021-03-04
Date modified: 2021-03-04

https://docs.cloudera.com/

Legal Notice

Contents

Using Streams Replication Manager in CDP Public Cloud overview

Streams Replication Manager (SRM) can be deployed in both CDP PvC Base (on-prem) and Data Hub (cloud) clusters. You can use your SRM deployment to replicate Kafka data between CDP PvC Base and Data Hub clusters, or to replicate data between multiple Data Hub clusters. Review the following information to learn more about your deployment options, as well as the prerequisites and use cases for using SRM in a cloud-based context.

Starting with the December 2020 release of CDP Public Cloud, SRM is included in the default Streams Messaging cluster definitions. As a result, you can deploy SRM in a Data Hub cluster and use it to replicate data between all types of CDP clusters. This includes deployments with either Data Hub, CDP PvC Base, or both.

The following sections provide information on how you can deploy SRM in a Data Hub cluster, what prerequisites you must meet before using SRM, and the common use cases where you can use SRM in a cloud-based context.

Differences between light and heavy deployments

In CDP Public Cloud, SRM can be deployed in Data Hub clusters with both the light and heavy duty variants of the Streams Messaging cluster definition. However, there are significant differences in how SRM is deployed with each definition:

Light duty definition:

In the light duty definition, SRM is deployed by default on the broker and master hosts of the cluster. This means that SRM is available for use by default in a Data Hub cluster provisioned with the light duty definition.

Heavy duty definition

In the heavy duty definition, SRM has its own host group. However, by default, the SRM host group is not provisioned. When creating a cluster with the heavy duty definition, you must set the instance count of the Srm nodes host group to at least one. Otherwise, SRM is not be deployed on the cluster.

For more information on cluster provisioning, see *Creating your first Streams Messaging cluster*. For more information on the default cluster definitions and cluster layouts, see *Streams Messaging cluster layout*.

Note: Deploying SRM in a Data Hub cluster requires version 7.2.6 or higher of Cloudera Runtime.

Prerequisites for using SRM

SRM can be used to replicate Kafka data between all types of CDP clusters. However, the following conditions must be met for all deployments and use cases:

- SRM must be able to access the Kafka hosts of the source and target cluster through the network.
- SRM must trust the TLS certificates of the brokers in the source and target clusters.

  This is required so that SRM can establish a trusted connection.
- SRM must have access to credentials that it can use to authenticate itself in both the source and target clusters.
- SRM must use a principal that is authorized to access Kafka resources (topics) on both source and target clusters.

Cloud-based use cases for SRM

There are three common uses cases when using SRM in a cloud-based context. These are as follows:

- Replicating data from a CDP PvC Base cluster to a Data Hub cluster with SRM running in the CDP PvC Base cluster.

  Note: Although in this use case SRM is not deployed in a Data Hub cluster, it is still considered as a cloud based use case as either the source or the destination of the replicated data is a Data Hub cluster.

- Replicating data from a CDP PvC Base cluster to a Data Hub cluster with SRM running in the Data Hub cluster.
- Replicating data between two Data Hub clusters.

Related Information

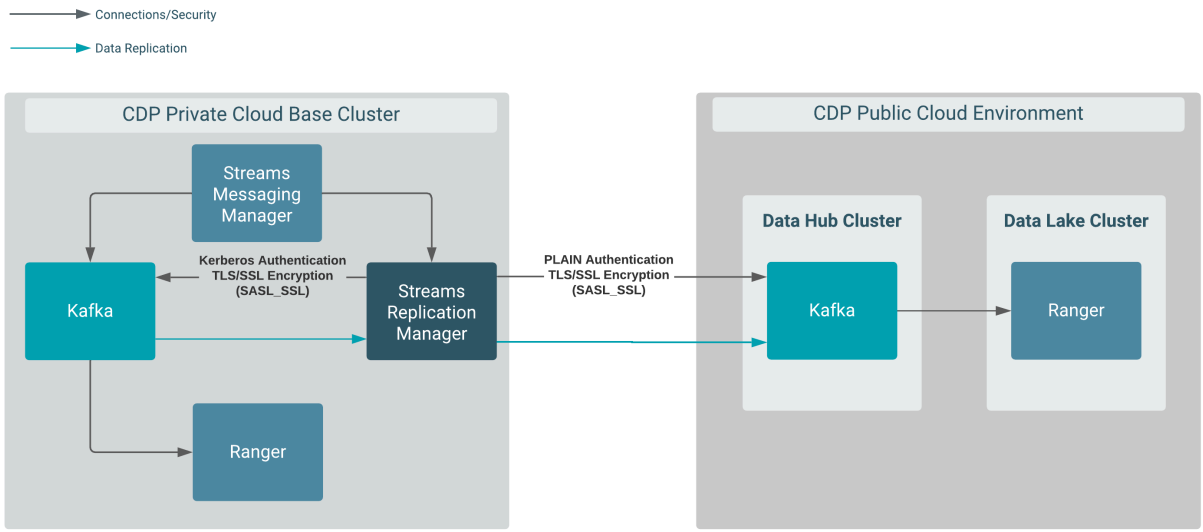
Creating your first Streams Messaging cluster
Streams Messaging cluster layout

Replicating data from CDP PvC Base cluster to Data Hub cluster with SRM running in CDP PvC Base cluster

You can set up and configure an instance of SRM running in a CDP PvC Base cluster to replicate data between the CDP PvC Base cluster and a Data Hub cluster. In addition, you can use SMM to monitor the replication process. Review the following example to learn how this can be set up.

About this task

Consider the following replication scenario:



In this scenario, data is replicated from a CDP PvC Base cluster that has Kafka, SRM, and SMM deployed on it. This is a secure cluster that has TLS/SSL encryption and Kerberos authentication enabled. In addition, it uses Ranger for authorization.

Data is being replicated from this cluster by SRM deployed in this cluster to a Data Hub cluster.

The Data Hub cluster is provisioned with the one of the default Streams Messaging cluster definitions.

Before you begin

This example scenario does not go into detail on how to set up the clusters and assumes the following:

- A Data Hub cluster provisioned with the Streams Messaging Light Duty or Heavy Duty cluster definition is available.

  For more information, see Creating your first Streams Messaging cluster in the CDF for Data Hub library. Alternatively, you can also review the cloud provider specific cluster creation instructions available in the Cloudera Data Hub library.

- A CDP PvC Base cluster with Kafka, SRM, and SMM is available. This cluster is TLS/SSL and Kerberos enabled. In addition, it uses Ranger for authorization.

  For more information, see the CDP Private Cloud Base Installation Guide.
- Network connectivity and DNS resolution are established between the clusters.

Important: In the following scenario, a new CDP machine user is created and set up specifically for SRM. Alternatively, it is also possible to use an existing machine user and skip steps 1 through 3, but this can only be done if the following requirements are met:

- The existing machine user has access to your CDP environment.
- The existing machine user has the correct Ranger permissions assigned to it.
- You have access to the existing machine user's credentials.

Procedure

1. Create a machine user for SRM in Management Console:

   A machine user is required so that SRM has credentials that it can use to connect to the Kafka service in the Data Hub cluster.

   a) Navigate to Management Console User Management.
   b) Click Actions Create Machine User.
   c) Enter a unique name for the user and click Create.

      For example: srm

      After the user is created, you are presented with a page that displays the user details.

      Note:

      The Workload User Name (srv_srm), is different from the actual machine user name (srm). The Workload User Name is the identifier you use to configure SRM.
   d) Click Set Workload Password.
   e) Type a password in the Password and Confirm Password fields. Leave the Environment field blank.
   f) Click Set Workload Password.
      A message appears on successful password creation.
2. Grant the machine user access to your environment:

   You must grant the machine user access to your environment for SRM to connect to the Kafka service with this user.

   a) Navigate to Management Console Environments, and select the environment where your Kafka cluster is located.
   b) Click Actions Manage Access.

      Use the search box to find and select the machine user you want to use.

      A list of **Resource Roles** appears.
   c) Select the EnvironmentUser role and click Update Roles.
   d) Go back to the **Environment Details** page and click Actions Synchronize Users to FreeIPA .
   e) On the Synchronize Users page, click Synchronize Users.

      Synchronizing users ensures that the role assignment is in effect for the environment.

      Important: Wait until this process is completed. Otherwise, you will not be able to continue with the next step.

3. Add Ranger permissions for the user you created for SRM in the Data Hub cluster:

   You must to grant the necessary privileges to the user so that the user can access Kafka resources. This is configured through Ranger policies.

   a) Navigate to Management Console Environments, and select the environment where your Kafka cluster is located.
   b) Click the Ranger link on the Environment Details page.
   c) Select the resource-based service corresponding to the Kafka resource in the Data Hub cluster.
   d) Add the Workload User Name of the user you created for SRM to the following Ranger policies:

      - All - consumergroup
      - All - topic
      - All - transactionalid
      - All - cluster
      - All - delegationtoken

4. Ensure that Ranger permissions exist for the streamsrepmgr user in the CDP PvC Base cluster:

   a) Access the Cloudera Manager instance of your CDP PvC Base cluster.
   b) Go to Ranger Ranger Admin Web UI.
   c) Log in to the Ranger Console (Ranger Admin Web UI).
   d) Ensure that the streamsrepmgr user is added to all required policies.

      If the user is missing, add it. The required policies are as follows:

      - All - consumergroup
      - All - topic
      - All - transactionalid
      - All - cluster
      - All - delegationtoken

5. Create a truststore on the CDP PvC Base cluster:

   A truststore is required so that the SRM instance running in the CDP PvC Base cluster can trust the secure Data Hub cluster. To do this, you extract the FreeIPA certificate from the CDP environment, create a truststore that includes the certificate, and copy the truststore to all hosts on the CDP PvC Base cluster.

   a) Navigate to Management Console Environments, and select the environment where your Kafka cluster is located.
   b) Go to the Summary tab.
   c) Scroll down to the FreeIPA section.
   d) Click ActionsGet FreeIPA Certificate.
      The FreeIPA certificate file, *[***ENVIRONMENT NAME***]*-env.crt, is downloaded to your computer.
   e) Run the following command to create the truststore:

      ```
      keytool \
        -importcert \
        -storetype JKS \
        -noprompt \
        -keystore datahub-truststore.jks \
        -storepass [***PASSWORD***] \
        -alias freeipa-ca \
        -file [***PATH TO FREEIPA CERTIFICATE***]
      ```

   f) Copy the datahub-truststore.jks file to a common location on all the hosts in your CDP PvC Base cluster.

      Cloudera recommends that you use the following location: /opt/cloudera/security/datahub-truststore.jks.
   g) Set the correct file permissions.

      Use 751 for the directory and 444 for the truststore file.

6. Configure the SRM properties in the CDP PvC Base cluster:
   a) Access the Cloudera Manager instance of your CDP PvC Base cluster.
   b) Go to Streams Replication Manager Configuration and configure the following properties:
      - Streams Replication Manager Cluster alias: datahub, cdppvc
      - Streams Replication Manager Driver Target Cluster: datahub, cdppvc

        Note: This property must either contain all aliases or left blank. Leaving the property blank has the same effect as adding all aliases.

      - Streams Replication Manager Service Target Cluster: datahub
      - Streams Replication Manager's Replication Configs:

        Important: Passwords that you set in this property are stored in plaintext.

```
#Bootstrap servers:
cdppvc.bootstrap.servers=[***MY-CDP-PVC-CLUSTER-
HOST-1.COM:9093***],[***MY-CDP-PVC-CLUSTER-HOST-2:9093***]
datahub.bootstrap.servers=[**MY-DATAHUB-CLUSTER-
HOST-1.COM:9093***],[***MY-DATAHUB-CLUSTER-HOST-1.COM:9093***]

#Replications:
cdppvc->datahub.enabled=true

#Security properties for the Datahub cluster:
datahub.security.protocol=SASL_SSL
datahub.sasl.mechanism=PLAIN
datahub.sasl.jaas.config=org.apache.kafka.common.security.plain.Plai
nLoginModule required username="[***WORKLOAD USER NAME***]" passwo
rd="[***MACHINE USER PASSWORD***]";
datahub.ssl.truststore.location=/OPT/CLOUDERA/SECURITY/DATAHUB-
TRUSTSTORE.JKS
datahub.ssl.truststore.password=[***PASSWORD***]

#Use the FQDN when specifying the cluster hosts.
#The terminating semicolon in the [***ALIAS***].sasl.jaas.config pr
operty must be included in the configuration.
#The value of the [***ALIAS***].ssl.truststore.location is the loca
tion where you copied the truststore in a previous step.
#The [***ALIAS***].ssl.truststore.password property must be specif
ied. Otherwise, the configuration might get overriden by the service
 ssl.truststore.password property.
```

        Note: This configuration does not specify the security properties required by SRM to connect to the co-located Kafka cluster (the cluster SRM is deployed in). This is because these properties are automatically passed by Cloudera Manager to SRM in the background.

   c) Click Save.
   d) Restart SRM.
   e) Deploy client configuration for SRM.
7. Start the replication process using the srm-control tool:
   a) SSH as an administrator to any of the SRM hosts in the CDP PvC cluster.

```
ssh [***USER***]@[***MY-CDP-PVC-CLUSTER.COM***]
```

   b) Create a configuration file for the srm-control tool.

      The srm-control tool behaves as a Kafka client and requires configuration that is similar to any Kafka client. The configuration file is specified with the --config option when you run the tool. The configuration file must include cluster alias definitions, as well as properties related to connection information and security. Cluster

aliases are defined a single time, connection and security properties are defined separately for each alias (cluster). In this example the file is named srm.properties.

```
#Define aliases:
clusters=datahub, cdppvc

#Bootstrap servers:
datahub.bootstrap.servers=[***MY-DATAHUB-CLUSTER-
HOST-1.COM:9093***],[***MY-DATAHUB-CLUSTER-HOST-1.COM:9093***]
cdppvc.bootstrap.servers=[***MY-CDP-PVC-CLUSTER-
HOST-1.COM:9093***],[***MY-CDP-PVC-CLUSTER-HOST-2:9093***]

#DataHub cluster's security properties:
datahub.security.protocol=SASL_SSL
datahub.sasl.mechanism=PLAIN
datahub.sasl.jaas.config=org.apache.kafka.common.security.plain.Pl
ainLoginModule required username="[***WORKLOAD USER NAME***]" passwo
rd="[***MACHINE USER PASSWORD***]";
datahub.ssl.truststore.location=/OPT/CLOUDERA/SECURITY/DATAHUB-
TRUSTSTORE.JKS
datahub.ssl.truststore.password=[***PASSWORD***]

#CDP PvC Base cluster's security properties:
cdppvc.security.protocol=SASL_SSL
cdppvc.sasl.mechanism=GSSAPI
cdppvc.sasl.kerberos.service.name=kafka
cdppvc.sasl.jaas.config=com.sun.security.auth.module.Krb5LoginModule r
equired useKeyTab=true keyTab="[***PATH TO KEYTAB FILE***]" storeKey=t
rue useTicketCache=false principal="[***MY KERBEROS PRINCIPAL****]";
cdppvc.ssl.truststore.location=[***CDP PVC BASE GLOBAL TRUSTSTORE
 LOCATION***]
cdppvc.ssl.truststore.password=[***CDP PVC BASE GLOBAL TRUSTSTORE
 PASSWORD***]

#Use the FQDN when specifying the cluster hosts.
#The terminating semicolon in the [***ALIAS***].sasl.jaas.config proper
ties must be included in the configuration.
#The value of the datahub.ssl.truststore.location property is the locati
on where you copied the truststore in a previous step.
```

c) Use the srm-control tool with the topics subcommand to add topics to the allow list:

```
srm-control --config ./SRM.PROPERTIES topics --source cdppvc --target
 datahub --add [***TOPIC NAME***]
```

d) Use the srm-control tool with the groups subcommand to add groups to the allow list:

```
srm-control --config ./SRM.PROPERTIES groups --source cdppvc --target
 datahub --add ".*"
```

8. Monitor replications.

Access the SMM UI in the CDP PvC Base cluster and go to the Cluster Replications page. The replications you set up will be visible on this page.

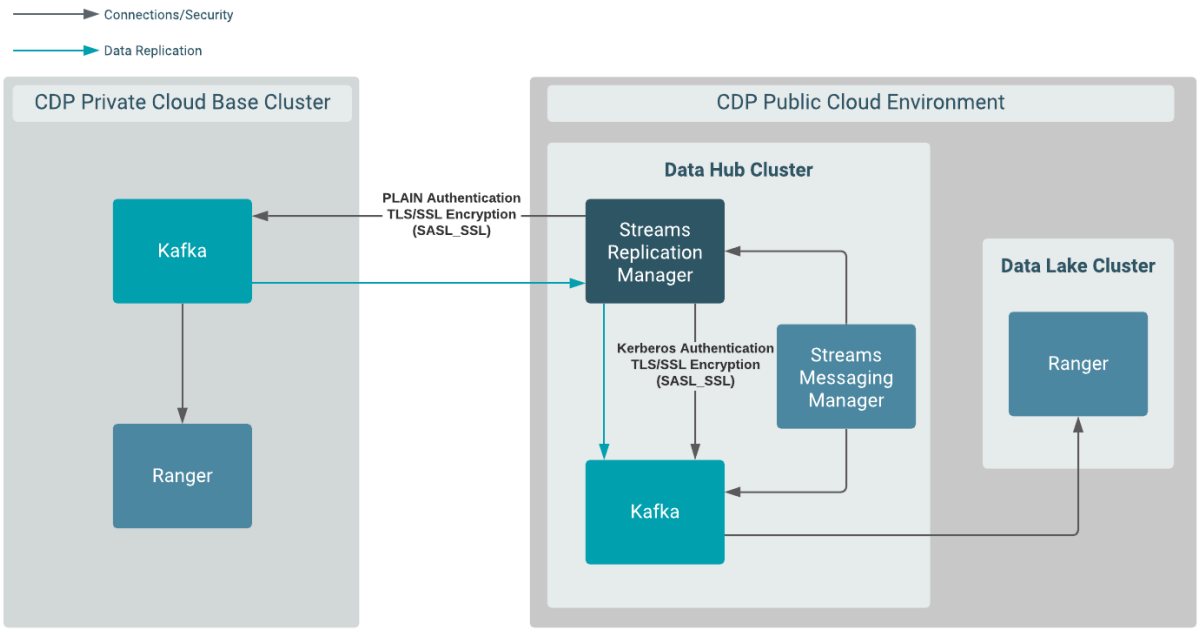
Note: If the topics or groups you added for replication are newly created, they might not be immediately visible. This is due to how frequently SRM checks for newly created topics and consumer groups. By default, this is set to 10 minutes, but can be configured with the Refresh Topics Interval Seconds and Refresh Groups Interval Seconds SRM properties. If at first your topics do not appear, wait a few minutes and refresh the page.

Replicating data from CDP PvC Base cluster to Data Hub cluster with SRM deployed in Data Hub cluster

You can set up and configure an instance of SRM running in a Data Hub cluster to replicate data between the Data Hub cluster and a CDP PvC Base cluster. In addition, you can use SMM to monitor the replication process. Review the following example to learn how this can be set up.

About this task

Consider the following replication scenario:



In this scenario, data is replicated from a CDP PvC Base cluster to a Data Hub cluster by an SRM instance that is deployed in the Data Hub cluster.

The CDP PvC Base cluster has Kafka deployed on it. It is a secure cluster that has TLS/SSL encryption enabled and uses PLAIN authentication. In addition, it uses Ranger for authorization.

The Data Hub cluster is provisioned with the one of the default Streams Messaging cluster definitions.

Before you begin

This example scenario does not go into detail on how to set up the clusters and assumes the following:

- A Data Hub cluster provisioned with the Streams Messaging Light Duty or Heavy Duty cluster definition is available.

  For more information, see Creating your first Streams Messaging cluster in the CDF for Data Hub library. Alternatively, you can also review the cloud provider specific cluster creation instructions available in the Cloudera Data Hub library.
- A CDP PvC Base cluster with Kafka is available. This cluster has TLS/SSL encryption enabled, uses PLAIN authentication, and has Ranger for authorization. For more information, see the CDP Private Cloud Base Installation Guide.
- Network connectivity and DNS resolution are established between the clusters.

Procedure

1. Obtain PLAIN credentials for SRM.

   The credentials of a PLAIN user that can access the CDP PvC Base cluster are required. These credentials are supplied to SRM in a later step. In this example *[***PLAIN USER***]* and *[***PLAIN USER PASSWORD***]* is used to refer to these credentials.

   Note: Typically, Kerberos cannot be easily configured to span across an on-premise and a public cloud environment. Therefore, these instructions assume that these environments use authentication methods that are easy to interoperate in such hybrid environments like LDAP or PAM authentication. For more information on how you can configure your Kafka service to use LDAP or PAM, see *Kafka Authentication*.

2. Add Ranger permissions for the PLAIN user in the CDP PvC cluster:

   You must ensure that the PLAIN user you obtained has correct permissions assigned to it in Ranger. Otherwise, SRM will not be able to access Kafka resources on the CDP PvC Base cluster.

   a) Access the Cloudera Manager instance of your CDP PvC Base cluster.
   b) Go to Ranger Ranger Admin Web UI.
   c) Log in to the Ranger Console (Ranger Admin Web UI).
   d) Add the *[***PLAIN USER***]* to the following policies:

   - All - consumergroup
   - All - topic
   - All - transactionalid
   - All - cluster
   - All - delegationtoken

3. Acquire the CDP PvC Base cluster truststore and add it to the Data Hub cluster:

   The actions you need to take differ depending on how TLS is set up in the CDP PvC Base cluster:

   For Auto TLS

   a. Obtain the certificate of the Cloudera Manager root Certificate Authority and its password.

   The Certificate Authority certificate and its password can be obtained using the Cloudera Manager API. The following steps describe how you can retrieve the certificate and password using the Cloudera Manager API Explorer. Alternatively, you can also retrieve the certificate and password by calling the appropriate endpoints in your browser window or using curl.

   1. Access the Cloudera Manager instance of your CDP PvC Base cluster.
   2. Go to SupportAPI Explorer.
   3. Find CertManagerResource.
   4. Select the /certs/truststore GET operation and click Try it out.
   5. Enter the truststore type.
   6. Click Execute.
   7. Click Download file under Responses.

      The downloaded file is your certificate.
   8. Select the /certs/truststorePassword GET operation and click Try it out.
   9. Click Execute.

      The password is displayed under Responses.

   b. Run the following command to create the truststore:

   ```
   keytool \
     -importcert \
     -storetype JKS \
     -noprompt \
     -keystore cdppvc-truststore.jks \
   ```

```
-storepass ***PASSWORD*** \
-alias cdppvc-cm-ca \
-file ***PATH TO CM CA CERTIFICATE***
```

Note down the password, it is needed in a later step.

c. Copy the cdpdc-truststore.jks file to a common location on all the hosts in your CDP Data Hub cluster.

Cloudera recommends that you use the following location: /opt/cloudera/security/cdppvc-truststore.jks.

d. Set the correct file permissions.

Use 751 for the directory and 444 for the truststore file.

For Manual TLS

a. Note down the CDP PvC Base cluster's truststore location and password, these should be known to you.

b. Copy the truststore file to a common location on all the hosts in your CDP Data Hub cluster.

Cloudera recommends that you use the following location: /opt/cloudera/security/truststore.jks.

c. Set the correct file permissions.

Use 751 for the directory and 444 for the truststore file.

4. Configure the SRM properties in the Data Hub cluster:

a) Access the Cloudera Manager instance of your Data Hub cluster.

b) Go to Streams Replication Manager Configuration and configure the following properties:

- Streams Replication Manager Cluster alias: datahub, cdppvc
- Streams Replication Manager Driver Target Cluster: datahub, cdppvc

Note: This property must either contain all aliases or left blank. Leaving the property blank has the same effect as adding all aliases.

- Streams Replication Manager Service Target Cluster: datahub
- Streams Replication Manager's Replication Configs:

Important: Passwords that you set in this property are stored in plaintext.

```
#Bootstrap servers:
cdppvc.bootstrap.servers=[***MY-CDP-PVC-CLUSTER-
HOST-1.COM:9093***],[***MY-CDP-PVC-CLUSTER-HOST-2:9093***]
datahub.bootstrap.servers=[***MY-DATAHUB-CLUSTER-
HOST-1.COM:9093***],[***MY-DATAHUB-CLUSTER-HOST-2.COM:9093***]

#Replications:
cdppvc->datahub.enabled=true

#Security properties for the CDP PvC Base cluster:
cdppvc.security.protocol=SASL_SSL
cdppvc.sasl.mechanism=PLAIN
cdppvc.sasl.jaas.config=org.apache.kafka.common.security.plain.P
lainLoginModule required username="[***PLAIN USER***]" passwo
rd="[***PLAIN USER PASSWORD***]";
cdppvc.ssl.truststore.location=/OPT/CLOUDERA/SECURITY/CDPPVC-
TRUSTSTORE.JKS
cdppvc.ssl.truststore.password=***PASSWORD***

#Use the FQDN when specifying cluster hosts.
#The terminating semicolon in the [***ALIAS***].sasl.jaas.config pr
operty must be included in the configuration.
#The value of the [***ALIAS***].ssl.truststore.location is the loca
tion where you copied the truststore in a previous step.
```

```
#The [***ALIAS***].ssl.truststore.password property must be specif
ied. Otherwise, the configuration might get overriden by the service
 ssl.truststore.password property.
```

Note: This configuration does not specify the security properties required by SRM to connect to the co-located Kafka cluster (the cluster SRM is deployed in). This is because these properties are automatically passed by Cloudera Manager to SRM in the background.

Note: The configuration of the security properties with the cdppvc. prefix depend on the security configuration of the CDP PvC Base cluster. This specific example is for a cluster that has TLS/SSL encryption and PLAIN authentication enabled. You need to change these configurations based on the setup of your CDP PvC Base cluster.

c) Click Save.
d) Restart SRM.
e) Deploy client configuration for SRM.

5. Start data replication topics using the srm-control tool:

a) SSH as an administrator to any of the SRM hosts in the Data Hub cluster.

```
ssh [***USER***]@[***MY-DATAHUB-CLUSTER.COM***]
```

b) Create a configuration file for the srm-control tool.

The srm-control tool behaves as a Kafka client and requires configuration that is similar to any Kafka client. The configuration file is specified with the --config option when you run the tool. The configuration file must include cluster alias definitions, as well as properties related to connection information and security. Cluster aliases are defined a single time, connection and security properties are defined separately for each alias (cluster). In this example the file is named srm.properties.

```
#Bootstrap servers:
cdppvc.bootstrap.servers=[***MY-CDP-PVC-CLUSTER-
HOST-1.COM:9093***],[***MY-CDP-PVC-CLUSTER-HOST-2:9093***]
datahub.bootstrap.servers=[***MY-DATAHUB-CLUSTER-
HOST-1.COM:9093***],[***MY-DATAHUB-CLUSTER-HOST-1.COM:9093***]

#CDP PVC Base cluster’s security properties:
cdppvc.security.protocol=SASL_SSL
cdppvc.sasl.mechanism=PLAIN
cdppvc.sasl.jaas.config=org.apache.kafka.common.security.plain.PlainLog
inModule required username="[***PLAIN USER***]" password="[***PLAIN USER
 PASSWORD***]";
cdppvc.ssl.truststore.location=/OPT/CLOUDERA/SECURITY/CDPPVC-
TRUSTSTORE.JKS
cdppvc.ssl.truststore.password=[***PASSWORD***]

#Data Hub cluster's security properties:
datahub.security.protocol=SASL_SSL
datahub.sasl.mechanism=GSSAPI
datahub.sasl.kerberos.service.name=kafka
datahub.sasl.jaas.config=com.sun.security.auth.module.Krb5LoginModule r
equired useKeyTab=true keyTab="[***PATH TO KEYTAB FILE***]" storeKey=t
rue useTicketCache=false principal="[***MY KERBEROS PRINCIPAL****]";
datahub.ssl.truststore.location=/OPT/CLOUDERA/SECURITY/DATAHUB-
TRUSTSTORE.JKS
datahub.ssl.truststore.password=[***PASSWORD***]

#Use the FQDN when specifying the cluster hosts.
#The terminating semicolon in the [***ALIAS***].sasl.jaas.config proper
ties must be included in the configuration.
```

```
#The value of the cdpdc.ssl.truststore.location property is the location
 where you copied the truststore in a previous step.
```

c) Use the srm-control tool with the topics subcommand to add topics to the allow list:

```
srm-control --config ./SRM.PROPERTIES topics --source cdppvc --target
 datahub --add [***TOPIC NAME***]
```

d) Use the srm-control tool with the groups subcommand to groups to the allow list:

```
srm-control --config ./SRM.PROPERTIES groups --source cdppvc --target
 datahub --add ".*"
```

6. Monitor the replication process.

Access the SMM UI in the Data Hub cluster and go to the Cluster Replications page. The replications you set up will be visible on this page.

Note: If the topics or groups you added for replication are newly created, they might not be immediately visible. This is due to how frequently SRM checks for newly created topics and consumer groups. By default, this is set to 10 minutes, but can be configured with the Refresh Topics Interval Seconds and Refresh Groups Interval Seconds SRM properties. If at first your topics do not appear, wait a few minutes and refresh the page.

Related Information

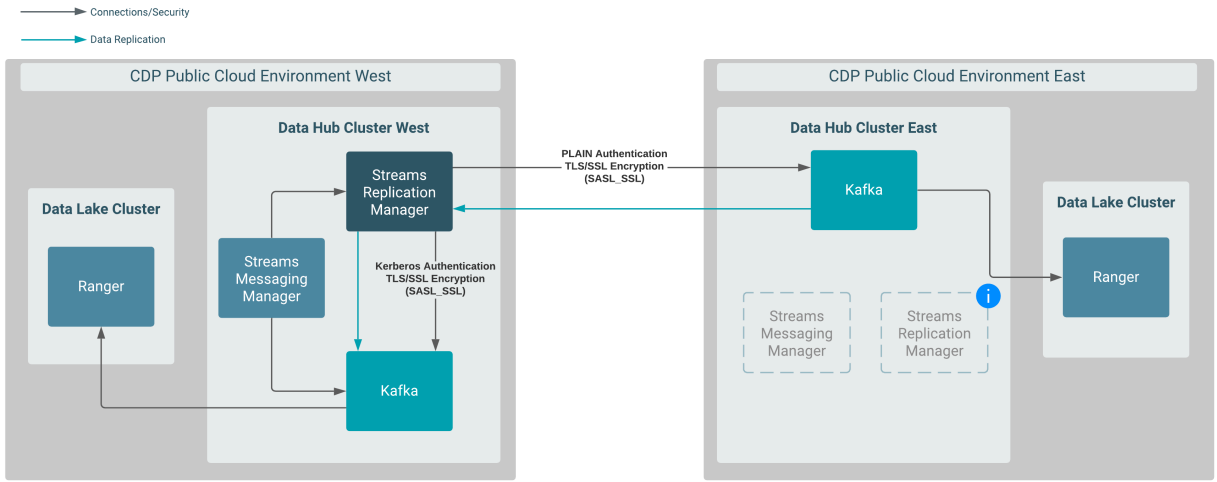
Kafka Authentication

Replicating data between Data Hub clusters with SRM deployed in a Data Hub cluster.

You can set up and configure an instance of SRM in a Data Hub cluster to replicate data between Data Hub clusters. In addition, you can use SMM to monitor the replication process. Review the following example to learn how this can be set up.

About this task

Consider the following replication scenario:



In this scenario, data is replicated between two Data Hub clusters that are provisioned in different CDP environments. More specifically, data in Data Hub East is replicated to Data Hub West by an instance of SRM running in Data Hub West.

Both Data Hub clusters are provisioned with the default Streams Messaging cluster definitions.

SRM and SMM are available in both clusters, but the instances in Data Hub East are not utilized in this scenario.

Before you begin

This example scenario does not go into detail on how to set up the clusters and assumes the following:

- Two Data Hub clusters provisioned with the Streams Messaging Light Duty or Heavy Duty cluster definition are available.

  For more information, see Creating your first Streams Messaging cluster in the CDF for Data Hub library. Alternatively, you can also review the cloud provider specific cluster creation instructions available in the Cloudera Data Hub library.
- Network connectivity and DNS resolution are established between the clusters.

Important: In the following scenario, a new CDP machine user is created and set up specifically for SRM. Alternatively, it is also possible to use an existing machine user and skip steps 1 through 3, but this can only be done if the following requirements are met:

- The existing machine user has access to your CDP environment.
- The existing machine user has the correct Ranger permissions assigned to it.
- You have access to the existing machine user's credentials.

Procedure

1. Create a machine user for SRM in Management Console:

   A machine user is required so that SRM has credentials that it can use to connect to the Kafka service in the Data Hub cluster. This step is only required in the environment where SRM is not running. In the case of this example, this is the CDP Public Cloud East environment.

   a) Navigate to Management Console User Management.
   b) Click Actions Create Machine User.
   c) Enter a unique name for the user and click Create.

      For example: srm

      After the user is created, you are presented with a page that displays the user details.

      Note:

      The Workload User Name (srv_srm), is different from the actual machine user name (srm). The Workload User Name is the identifier you use to configure SRM.
   d) Click Set Workload Password.
   e) Type a password in the Password and Confirm Password fields. Leave the Environment field blank.
   f) Click Set Workload Password.
      A message appears on successful password creation.

2. Grant the machine user access to your environment:

   You must to grant the machine user access in your environments, otherwise SRM will not be able to connect to the Kafka service with this user. This step is only required in the environments where SRM is not running. In the case of this example this is the CDP Public Cloud East environment.

   a) Navigate to Management Console Environments, and select the environment where your Kafka cluster is located.
   b) Click Actions Manage Access.

      Use the search box to find and select the machine user you want to use.

      A list of **Resource Roles** appears.
   c) Select the EnvironmentUser role and click Update Roles.
   d) Go back to the **Environment Details** page and click Actions Synchronize Users to FreeIPA .
   e) On the Synchronize Users page, click Synchronize Users.

      Synchronizing users ensures that the role assignment is in effect for the environment.

      Important: Wait until this process is completed. Otherwise, you will not be able to continue with the next step.

3. Add Ranger permissions for the user you created for SRM.

   This step is only required in the environment where SRM is not running. In the case of this example the environment is the CDP Public Cloud East .

   a) Navigate to Management Console Environments, and select the environment where your Kafka cluster is located.
   b) Click the Ranger link on the Environment Details page.
   c) Select the resource-based service corresponding to the Kafka resource in the Data Hub cluster.
   d) Add the Workload User Name of the user you created for SRM to the following Ranger policies:

      - All - consumergroup
      - All - topic
      - All - transactionalid
      - All - cluster
      - All - delegationtoken

4. Establish trust between the clusters:

   A truststore is needed so that the SRM instance running in Data Hub West can trust Data Hub East. To do this, you extract the FreeIPA certificate from Environment East, create a truststore that includes the certificate, and copy the truststore to all hosts on Data Hub West.

   a) Navigate to Management Console Environments, and select Environment East.
   b) Go to the Summary tab.
   c) Scroll down to the FreeIPA section.
   d) Click ActionsGet FreeIPA Certificate.
      The FreeIPA certificate file, *[***ENVIRONMENT NAME***]*-env.crt, is downloaded to your computer.
   e) Run the following command to create the truststore:

```
keytool \
  -importcert \
  -storetype JKS \
  -noprompt \
  -keystore truststore-east.jks \
  -storepass [***PASSWORD***] \
  -alias freeipa-east-ca \
  -file [***PATH TO FREEIPA CERTIFICATE***]
```

   f) Copy the truststore-east.jks file to a common location on all the hosts in your Data Hub West cluster.

      Cloudera recommends that you use the following location: /opt/cloudera/security/truststore-east.jks.

g) Set the correct file permissions.

Use 751 for the directory and 444 for the truststore file.

5. Configure the SRM properties in the Data Hub West cluster:

a) Access the Cloudera Manager instance of the Data Hub West cluster.

b) Go to Streams Replication Manager Configuration and configure the following properties:

- Streams Replication Manager Cluster alias: dheast, dhwest
- Streams Replication Manager Driver Target Cluster: dheast, dhwest

Note: This property must either contain all aliases or left blank. Leaving the property blank has the same effect as adding all aliases.

- Streams Replication Manager Service Target Cluster: dhwest
- Streams Replication Manager's Replication Configs:

Important: Passwords that you set in this property are stored in plaintext.

```
#Bootstrap servers:
dheast.bootstrap.servers=[***MY-DATAHUB-EAST-CLUSTER-
HOST-1.COM:9093***],[***MY-DATAHUB-EAST-CLUSTER-HOST-2:9093***]
dhwest.bootstrap.servers=[***MY-DATAHUB-WEST-CLUSTER-
HOST-1.COM:9093***],[***MY-DATAHUB-WEST-CLUSTER-HOST-1.COM:9093***]

Replications:
dheast->dhwest.enabled=true
#Datahub East cluster’s security properties:
dheast.security.protocol=SASL_SSL
dheast.sasl.mechanism=PLAIN
dheast.sasl.jaas.config=org.apache.kafka.common.security.plain.Plain
LoginModule required username="[***WORKLOAD USER NAME***]" passwo
rd="[***MACHINE USER PASSWORD***]";
dheast.ssl.truststore.location=/OPT/CLOUDERA/SECURITY/TRUSTSTORE-
EAST.JKS
dheast.ssl.truststore.password=[***PASSWORD***]

#Use the FQDN when specifying cluster hosts.
#The terminating semicolon in the [***ALIAS***].sasl.jaas.config pr
operty must be included in the configuration.
#The value of the [***ALIAS***].ssl.truststore.location is the loca
tion where you copied the truststore in a previous step.
#The [***ALIAS***].ssl.truststore.password property must be specif
ied. Otherwise, the configuration might get overriden by the service
 ssl.truststore.password property.
```

Note: This configuration does not specify the security properties required by SRM to connect to the co-located Kafka cluster (the cluster SRM is deployed in). This is because these properties are automatically passed by Cloudera Manager to SRM in the background.

c) Click Save.

d) Restart SRM.

e) Deploy client configuration for SRM.

f) Restart SMM.

6. Start data replication topics using the srm-controltool:
   a) SSH as an administrator to any of the SRM hosts in the Data Hub West cluster.

```
ssh [***USER***]@[***DATA-HUB-WEST-CLUSTER-HOST-1.COM***]
```

   b) Create a configuration file for the srm-control tool.

      The srm-control tool behaves as a Kafka client and requires configuration that is similar to any Kafka client. The configuration file is specified with the --config option when you run the tool. The configuration file must include cluster alias definitions, as well as properties related to connection information and security. Cluster aliases are defined a single time, connection and security properties are defined separately for each alias (cluster). In this example the file is named srm.properties.

```
Define aliases:
clusters=dheast, dhwest
#Bootstrap servers
dheast.bootstrap.servers=[***MY-DATAHUB-EAST-CLUSTER-
HOST-1.COM:9093***],[***MY-DATAHUB-EAST-CLUSTER-HOST-2:9093***]
dhwest.bootstrap.servers=[***MY-DATAHUB-WEST-CLUSTER-
HOST-1.COM:9093***],[***MY-DATAHUB-WEST-CLUSTER-HOST-1.COM:9093***]

#Datahub East cluster’s security properties:
dheast.security.protocol=SASL_SSL
dheast.sasl.mechanism=PLAIN
dheast.sasl.jaas.config=org.apache.kafka.common.security.plain.Plai
nLoginModule required username="[***WORKLOAD USER NAME***]" passwo
rd="[***MACHINE USER PASSWORD***]";
dheast.ssl.truststore.location=/OPT/CLOUDERA/SECURITY/TRUSTSTORE-
EAST.JKS
dheast.ssl.truststore.password=[***PASSWORD***]

#Datahub West cluster’s security properties:
dhwest.security.protocol=SASL_SSL
dhwest.sasl.mechanism=GSSAPI
dhwest.sasl.kerberos.service.name=kafka
dhwest.sasl.jaas.config=com.sun.security.auth.module.Krb5LoginModule r
equired useKeyTab=true keyTab="[***PATH TO KEYTAB FILE***]" storeKey=t
rue useTicketCache=false principal="[***MY KERBEROS PRINCIPAL****]";
dhwest.ssl.truststore.location=/OPT/CLOUDERA/SECURITY/TRUSTSTORE-
WEST.JKS
dhwest.ssl.truststore.password=[***PASSWORD***]

#Use the FQDN when specifying the cluster hosts.
#The terminating semicolon in the [***ALIAS***].sasl.jaas.config proper
ties must be included in the configuration.
#The value of the dheast.ssl.truststore.location property is the locatio
n where you copied the truststore in a previous step.
```

   c) Use the srm-control tool with the topics subcommand to add topics to the allow list:

```
srm-control --config ./SRM.PROPERTIES topics --source dheast --target
 dhwest --add [***TOPIC NAME***]
```

   d) Use the srm-control tool with the groups subcommand to add groups to the allow list:

```
srm-control --config ./SRM.PROPERTIES groups --source dheast --target
 dhwest --add ".*"
```

7. Monitor the replication process.

   Access the SMM UI in the Data Hub West cluster and go to the Cluster Replications page. The replications you set up will be visible on this page.

   Note: If the topics or groups you added for replication are newly created, they might not be immediately visible. This is due to how frequently SRM checks for newly created topics and consumer groups. By default, this is set to 10 minutes, but can be configured with the Refresh Topics Interval Seconds and Refresh Groups Interval Seconds SRM properties. If at first your topics do not appear, wait a few minutes and refresh the page.