Cloudera Public Cloud

Data Migration Tools and Methods

Date published: 2019-08-22
Date modified: 2025-08-18

CLOUDERA

Legal Notice

Contents

Data Migration Tools and Methods Overview

Describes the Cloudera data migration services that help you to understand and optimize your existing workloads, clusters, and resources, plan your migration, and migrate your workload data.

Related Information
Accelerate Your Migration to Cloudera with Workload Manager or Workload XM
Use Cloudera Replication Manager to migrate to Cloudera on cloud

Accelerate Your Migration to Cloudera with Workload Manager or Workload XM

Overview of how Workload Manager and Workload XM can accelerate your migration from legacy CDH and HDP to Cloudera by reducing migration risks.

As part of your migration best practices, it is important to perform an in-depth analysis of your workloads before migrating to Cloudera. Workload Manager or Workload XM enable you to interactively understand your existing workloads, clusters, and resources. The wide range of metrics (collected by Telemetry Publisher or Databus Producer) and health tests help you identify and troubleshoot issues and the prescriptive guidance and recommendations help you quickly address and optimize those problems before or after migration. Finally, the performance baseline feature and historical analysis helps you identify and address performance problems and compare a workload's performance before and after your migration.

The following steps describe the tasks that you perform to plan and accelerate your migration to Cloudera. These steps provide you with an in-depth knowledge to make migration decisions by enabling you to identify which workloads are good candidates for cloud migration, plan which workloads, queries, and jobs require optimization and whether optimization is performed before or after migration, and identify and address performance problems by establishing baselines from health issues that will also enable a performance comparison of your workloads from before and after your migration:

- Identify current and potential issues and the priority in which your workloads are migrated from your legacy platform to a Cloudera form factor. This includes:

    - Monitoring and identifying issues in your existing clusters and recording their historical analysis.
    - Analyzing your existing resource consumption and resource contentions.
    - Identifying rogue users and resource hungry workloads.
- Plan which workloads, jobs, and queries require optimization and whether optimization is performed before or after migration.
- Enable a before and after migration workload performance comparison by capturing performance baselines.

Workload Manager and Workload XM

Workload Analysis Steps

1

Step 1 Identify Current and Potential Issues

Identify current and potential issues and the priority in which your workloads are migrated from your legacy platform to a CDP form factor.

2

Step 2 Create an Optimization Plan

Plan which workloads, jobs, and queries require optimization and whether optimization is performed before or after migration.

3

Step 3 Capture Your Existing Baselines

Enable a before and after migration workload performance comparison by capturing performance baselines.

Step 1 Identify Current and Potential Issues

Describes the tasks that help you discover which workloads to migrate and which workloads and users are currently containing issues and impacting costs, such as those resources that are consuming large amounts of CPU and memory.

Performing an in-depth analysis of your workloads helps you understand the level of effort required and helps you schedule and prioritize the order in which your workloads are migrated. Workload Manager and Workload XM identify existing and potential problems and their prescriptive guidance and recommendations enable you to quickly address and optimize those problems before or after migration.

The following tasks help you identify and analyze the state of your current workloads.

Identifying Workload Problems and Health Issues

This task identifies what workloads are running on your CDH and HDP clusters and if there are any health issues that exist or were encountered and not addressed.

About this task

Describes how to use Workload Manager or Workload XM to quickly identify resource and workload health issues.

Note: The health test metrics provided by Workload Manager and Workload XM are dependent on the workload's processing engine.

The following chart widgets quickly enable you to locate what engines that are running on your CDH and HDP clusters and what jobs and queries are failing the health tests:

- The cluster's Usage Analysis chart widget enables you to visually see what engines are running on the cluster, how many jobs or queries are processed by the engine, and how many jobs or queries have failed and missed their SLA.
- The Suboptimal chart widget enables you to visually see at a glance what issues are currently impacting your jobs or queries and how they are executing on your cluster.

  This chart uses the Workload Manager and Workload XM health tests and shows the distribution of jobs and queries that failed. The health tests are performed when a job or query has finished and provide insight into the performance of the job or query, such as how much data was processed and how long it took to process.

  The Health Check list, on the engine's Jobs or Queries page, categorizes the health tests. For example, for an Impala engine, the Aggregation Spilled Partitions, HashJoin Spilled Partitions, and Slow Client health checks, test for resource health issues. For Hive, MapReduce, Oozie, and Spark engines, the Insufficient Partitioning and Many Materialized Columns health checks, test for query and schema issues, such as, is the code using SELECT * on millions of columns.

  These categories further enable you to understand the type of problem:

  - Metadata/Statistics
  - Optimal Configuration
  - Performance
  - Query/Schema Design
  - Skew
  - Baseline

For more information about each health check, see *RELATED INFORMATION*.

Procedure

1. In a supported browser, log in to the Workload Manager or the Workload XM web UI.
2. In the Clusters page, select the cluster required for analysis.
3. From the Usage Analysis chart widget, record the engines running workloads and note any issues.
4. From the Navigation panel or the Usage Analysis chart widget, click an engine that requires more analysis.

5. Scroll down to the engine's Suboptimal widget chart, as shown in the following example image. Record the health tests that contain a bar, which denote health tests that were violated.
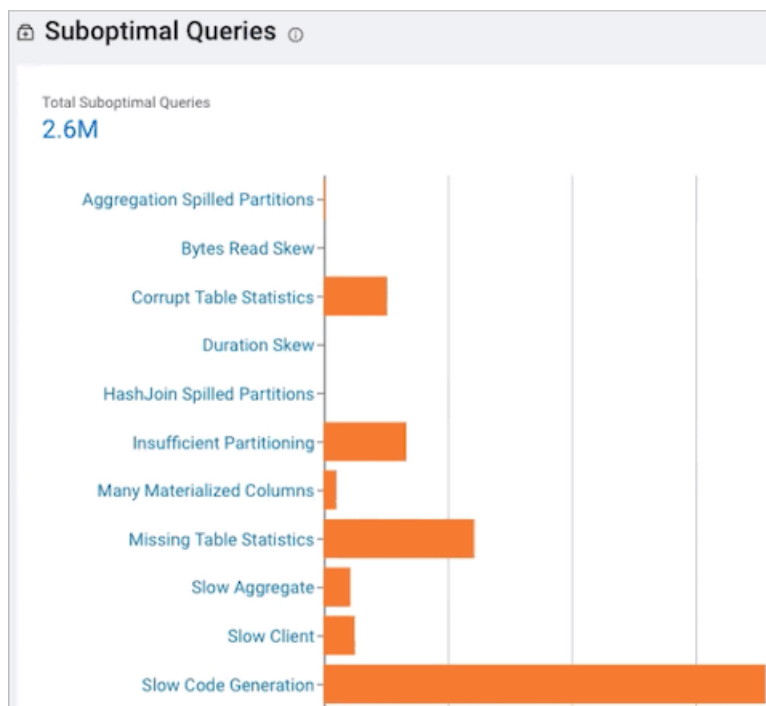
6. For more insights, click on any bar that displays a failed health check.

The Jobs or Queries page opens.

7. Click a job or query with a Health Issue and then select the Health Checks tab.

The Health Checks page is divided into three sections that display information about the job or query. Where:

- The left section lists the health checks that were performed.
- The middle section displays the stages of the job or query where a health check was performed. By default, all healthy stages are hidden and are revealed when a stage is selected.
- The right section lists detailed information about a stage. Failed stages list the failed health check, the diagnosis, and recommendations.

8. Either record the diagnosis and prescriptive recommendation or use the recommendation to fix the issue.

Related Information

Hive, MapReduce, Oozie, and Spark Health Checks

Impala Health Checks

Identifying Resource Contention

This task identifies the resource usage and consumption of your workloads.

About this task

Describes how to use Workload Manager or Workload XM to identify workload resource consumption and contention problems by observing the types of resources and the amount of memory that is consumed.

Workload Manager and Workload XM provide the following chart widgets that help you analyze and identify resource consumption and contention problems:

- The Resource Consumption By Services chart widget, which is displayed in the cluster Summary page. It shows the CPU and memory consumption for each service across the time range you selected. Hover your mouse over the time line, to display the amount of CPU or memory, as a percentage, that is consumed by each of the cluster's services.

- The Resource Consumption By Nodes chart widget, which is displayed in the cluster Summary page. It shows the CPU and memory consumption for each node in the cluster. Hover your mouse over the time line, to display the amount of CPU or memory, as a percentage, that is consumed by each node and its services.
- The Memory Utilization chart widget, which is displayed in either the Impala engine or Workloads (Workload View) page. It shows the aggregated maximum amount of memory that is used by the queries on any node performing the processing. It helps you identify inefficient queries that are consuming the most amount of memory during processing and if you need to allocate more memory to continue running your queries.
- The Resource Consumption chart widget, which is displayed in either the Impala engine or Workloads (Workload View) page. It shows the concurrent use of CPU and memory consumption for a workload across the time line you selected.

> ⚠️ **Important:** This task assumes that you have created workload views. For more information on how to create a Workload View, see *RELATED INFORMATION*.

Procedure

1. In a supported browser, log in to the Workload Manager or the Workload XM web UI.
2. In the Clusters page, select the cluster required for analysis.
3. From the Navigation panel, select Workloads.
4. In the Workloads page, select the required Workload.
5. Locate the Resource Consumption chart widget and identify any resource contentions by comparing the CPU Core Hours with Memory Usage.
6. Record any above average CPU and memory usage.

   For example, the following Resource Consumption chart widget image shows spikes when CPU and memory usage is high. This could be due to several job events occurring at the same time. The memory usage drops significantly around May 10th, which will need further investigation.
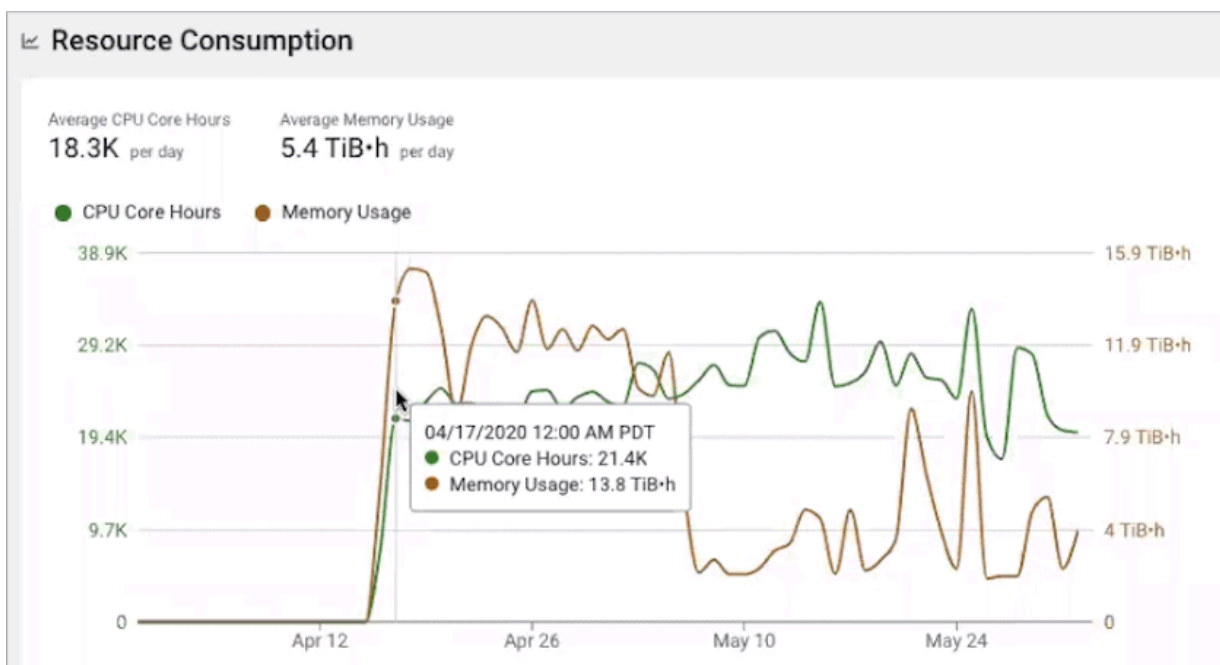
To rule out any serious problems, Cloudera recommends monitoring your memory and CPU usage from one of the aforementioned resource chart widgets. As part of your migration plan, Cloudera recommends eliminating any resource contention before you migrate. Otherwise, these problems will resurface in Cloudera.

Related Information

Classifying Workloads for Analysis with Workload Views

Identifying Rogue Users from a Workload View

This task identifies rogue users, which are users that consume excessive amounts of resources that can impact your costs.

About this task

Describes how to use Workload Manager or Workload XM to identify rogue users.

> ⚠️ **Important:** This task assumes that you have created workload views. For more information on how to create a Workload View, see *RELATED INFORMATION*.

Procedure

1. In a supported browser, log in to the Workload Manager or the Workload XM web UI.
2. In the Clusters page, select the cluster required for analysis.
3. From the Usage Analysis chart widget, select the required engine, such as Impala.
4. In the engine page, locate the engine Usage Analysis chart widget.
5. From the Top Users by list do the following:
   a. Select the # Queries option and identify any rogue users who are running more than the average share of jobs or queries.
   b. Select the Memory option and identify any rogue users whose jobs or queries are consuming the largest amount of memory.
   c. Select the CPU option and identify any rogue users whose jobs or queries are over exceeding the cluster's CPU.

6. Record any problems.

For example, when comparing memory consumption with the number of queries assigned to all users in the following Usage Analysis chart widget you will notice that the 3xksle8z user is running only 3% of the queries assigned, yet their queries are consuming more memory (5.9 PiB) than any other user.

Usage Analysis

Users   Pools   Databases

Top Users by: # Queries ∨

User	# Queries	%
zxLKfheu1	175.8K	67%
fkjh3fk45	24.7K	9%
9f8h3dd6	14.4K	6%
zv3dckh4	12K	5%
w38hckj3	7.8K	3%
3xksle8z	7.4K	3%

Usage Analysis

Users   Pools   Databases

Top Users by: Memory ∨

User	Memory
3xksle8z	5.9 PiB × s
fkjh3fk45	2.9 PiB × s
9f8h3dd6	355.4 TiB × s
zv3dckh4	87.6 TiB × s
w38hckj3	82.6 TiB × s
zxLKfheu1	68 TiB × s
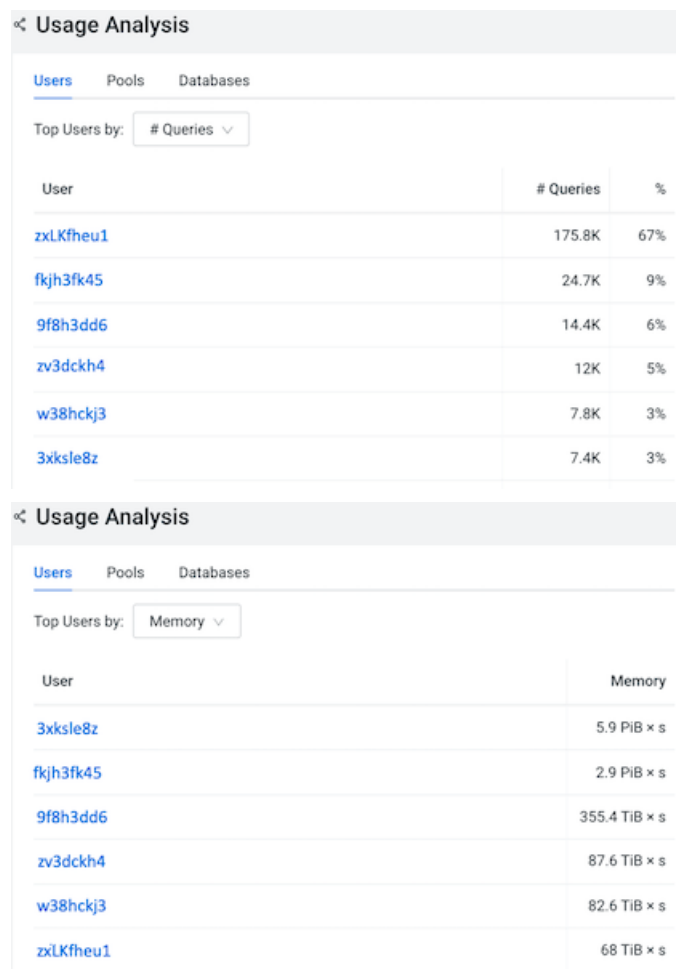
To discover which queries executed by user 3xksle8z are consuming the largest amount of memory, a new workload view was created to track this user's resource consumption.

The following image shows the metrics of a SQL Statement that was identified as using an excessive amount of CPU from the new workload view, where there was over 200hrs CPU usage during a 34 hour period. Further

metrics are displayed that show performance issues, identified as aggregation issues, and that show there are 14 tables missing table statistics.
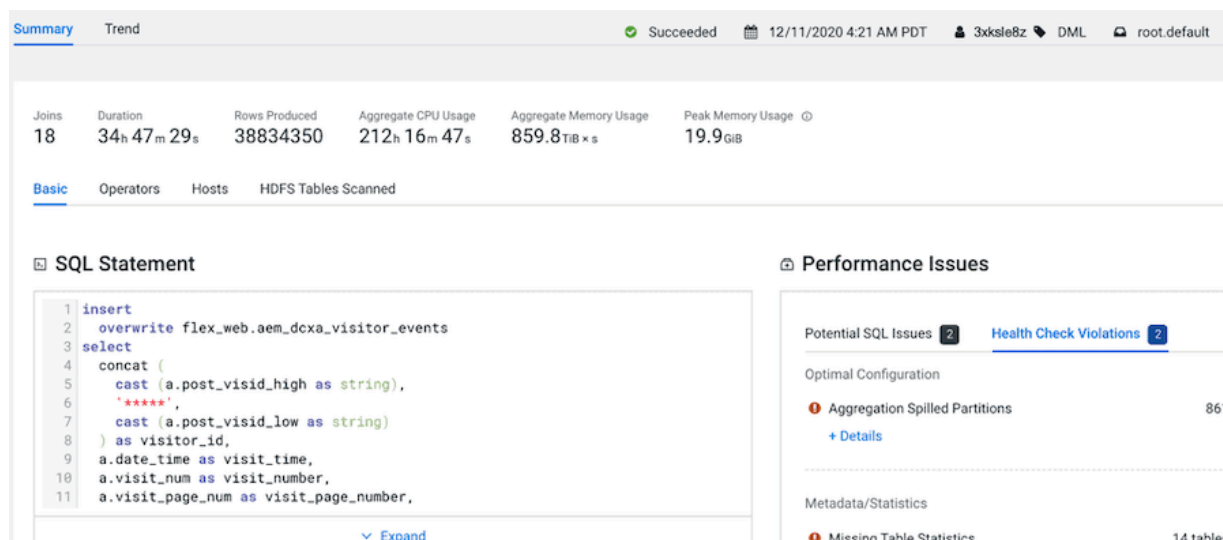
Related Information
Classifying Workloads for Analysis with Workload Views

Identifying Resource-Hungry Workloads

This task compares a workload with all other workloads in your cluster for identifying workloads that are using an excessive amount of resources.

Workload Manager and Workload XM enables you to compare the resource consumption of a particular workload or user against all other workloads or users in the cluster.

⚠️ **Important:** This task assumes that you have created workload views. For more information on how to create a Workload View, see *RELATED INFORMATION*.

For example, you can compare the resource consumption of a specific workload against all the workloads in the cluster with the Resource Consumption chart widget, as shown in the following image:

When you compare the two chart metrics, the chart on the left shows that of a single workload that is consuming half the memory; 1.9 TiB verses 3.7 TiB.

You can delve deeper by creating a Workload View whose user criteria value is the user whose jobs are consuming an inordinate amount of memory and then create a Workload View whose user criteria value comprises all the users except the user that is of interest.

When you compare the two chart metrics over the same time frame, the chart on the left shows the memory consumption as 84 GiB and the CPU core hours as 6.1K compared to 3.7 TiB and 20K.

You could migrate this user to Cloudera and relieve some resources from your on-premises system.

Related Information

Classifying Workloads for Analysis with Workload Views

Step 2 Create an Optimization Plan

Describes how an Optimization Plan can accelerate your migration by defining which workloads, jobs, and queries to optimize before and which workloads to optimize after migration.

Your optimization plan should include a list of workloads or set of workloads and their existing status, including any current performance issues and health check violations, and the suggested improvement recommendations by Workload Manager or Workload XM. From your list of insights, decide which workloads, jobs, and queries require optimization before migration and which workloads, jobs, and queries require optimization after migration.

Your optimization plan should also consider:

• The type and severity of the issue
• The Cloudera form factor
• Your data infrastructure
• The number of clusters in your Cloudera environment
• Memory sizing considerations
• The before and after features, engines, and services

For example, during the processing of a SQL statement, the performance health tests discovered aggregate issues and missing table statistics.
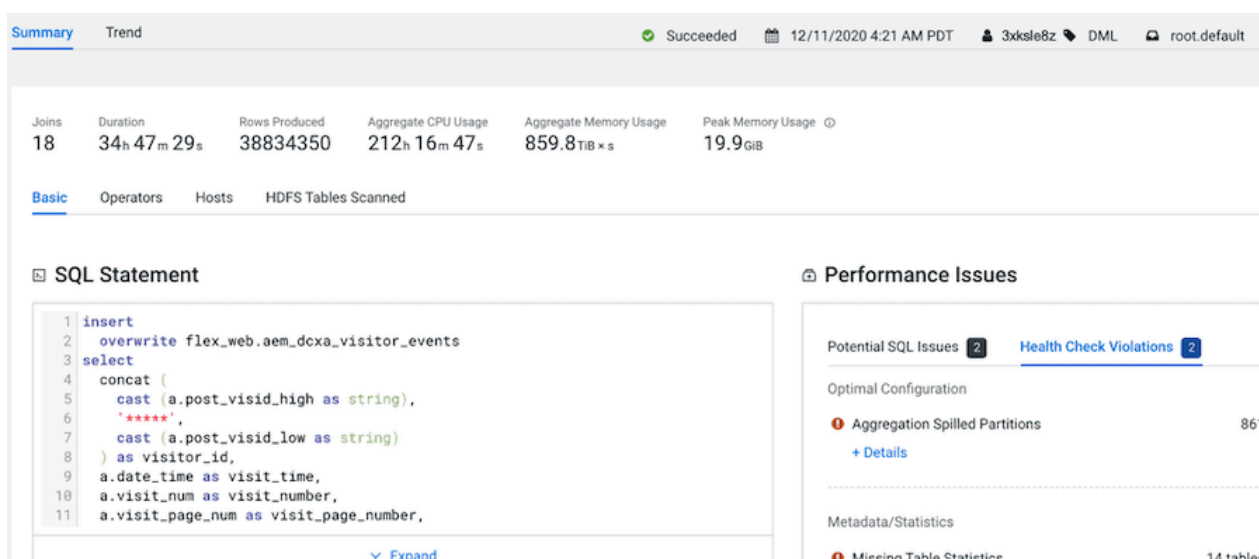
When a performance issue lists Aggregate Spilled Partitions as the cause it denotes memory issues. To ensure that all your SQL queries have sufficient memory when processing you should consider increasing your cluster's memory size. For this issue, the optimization would be solved as part of your Cloudera infrastructure and sizing considerations. Where you would use this information to research the costs of more memory for your Cloudera clusters as this would ensure enough memory for other such queries.

For the missing table statistics issue, it would also be more productive to collect the table statistic metadata after migration. Therefore, for this issue the optimization should be performed after migration.

Other considerations include:

- Understanding what features and engines are available in Cloudera.

  For example, users will have the ability to use Hive on TEZ and Hive LLAP in Cloudera, which are not available in CDH or HDP. Therefore, consider optimizing your Hive engine issues after migration.
- Your legacy applications and workloads that are at the end of their life cycle or those that would require development in order to work in Cloudera.
- Any issue or condition that would result in migrating garbage data. For example, you can immediately improve the processing performance of an Impala query by rewriting any poor SQL code, as described in the next topic *IDENTIFYING AND CORRECTING INEFFICIENT SQL CODE*.
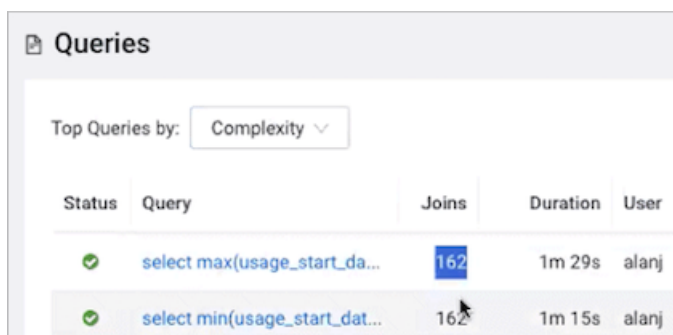
Identifying and Correcting Inefficient SQL Code

Cloudera Observability dissects inefficient Impala SQL code and provides prescriptive tuning recommendations.

You can immediately improve the processing performance of an inefficient Impala query by rewriting any poor SQL code. For example, to prevent skewness, SQL operations that use JOINS clauses may require changes to the SQL code by selecting columns with the most evenly distributed values. Or, as in the following example, having too many joins and inline views are characteristic of inefficiently written SQL code.

This example uses the Queries chart widget, which provides the Duration, Complexity, CPU, and Memory options that help you identify SQL problems.
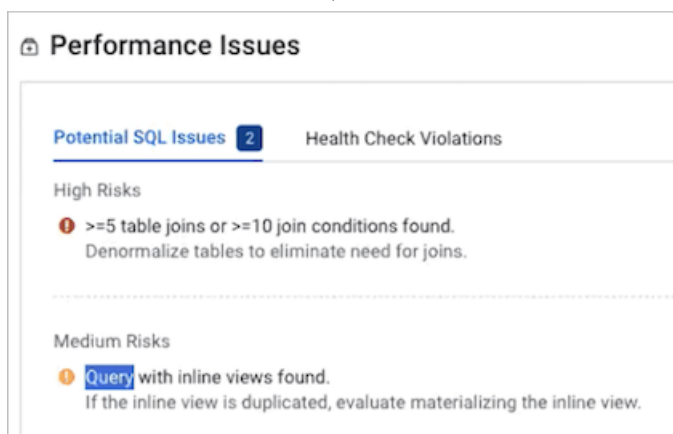
In this Impala engine case, several queries displayed a larger than average duration time in the Duration option of the Queries chart widget. When the Complexity option was selected it displayed two queries that contain a large number of joins, as shown in the following image:

Even though Impala can process hundreds of joins in a minute, reducing an inordinate number of joins in an SQL statement will improve performance.

When we drill down further in Cloudera Observability by selecting one of the queries of interest, the Performance Issues, which are related to the SQL statement, display the Cloudera Observability prescriptive tuning recommendations. In this case, the recommendation is to denormalize the tables and materialize the inline views:

Step 3 Capture Your Existing Baselines

Describes how to create baselines that enable you to address performance problems and compare the performance of your workloads, jobs, and queries before and after migration.

Workload Manager and Workload XM baseline metrics measure the current performance of a job against the average performance of previous runs. They use performance data from 30 of the most recent runs of a job and require a minimum of three runs. The baseline comparisons start with the fourth run of a job.

Note: When a baseline is first created there will be comparison differences until more data is established.

Cloudera highly recommends that before you migrate, you establish and capture your workloads, jobs, and query baselines with Workload Manager or Workload XM. Baselines enable you to identify and address performance problems, as well as, enable you to compare the performances of your workloads, jobs, and queries before and after migration.

For more information on how to establish and display baselines and create job comparisons, see *RELATED INFORMATION*.

The following images show some of the baseline metrics that Workload Manager and Workload XM provide:

- This image shows a few of the performance metrics listed in the Baseline tab:

- This image shows the comparison between the baseline performance metrics and the current job run:

To capture a job's baseline performance:

1. In a supported browser, log in to the Workload Manager or the Workload XM web UI.
2. In the Clusters page, select the cluster required for analysis.
3. From the Usage Analysis chart widget, select the required engine, such as Spark.
4. From the Slow Jobs chart widget, click on a job and record its baseline and any prescriptive improvements that can be made.
5. If a baseline is not displaying, click Compare with Previous Run.
6. Capture the Job Comparison page and record any health issues.

You can also compare two different runs of the same job with the Job comparison feature.

To compare two different runs of the same job:

1. In a supported browser, log in to the Workload Manager or the Workload XM web UI.
2. In the Clusters page, select the cluster required for analysis.
3. In the Trend chart widget, select the tab of an engine whose jobs you want to analyze and then click its Total Jobs value.

   The engine's Jobs page opens.
4. List and display details of all the runs of a specific job by selecting one of the job runs and then, in the Jobs details page, click the Trends tab.

5. To compare two job runs, select the check boxes adjacent to the job runs you require and then click Compare.

   The Job Comparison page opens displaying more details about each job.

Before moving to Cloudera, it is highly recommended capturing those jobs that have a maximum impact on your CDH or HDP clusters and establishing their baselines. After migrating to Cloudera, capture those job baselines again with Workload Manager or Workload XM. If you see significant deviation between the CDH or HDP and Cloudera baselines, drill down further in Workload Manager or Workload XM to understand the effects of the migration.

You can also identify trends as well as baselines by analyzing your engine's or cluster's performance trends from the Trends chart widget and the Trend tab. As described in the next topic.

Related Information
Troubleshooting with the Job Comparison Feature

Identifying Performance Trends

This task identifies performance trends over a selected time range.

About this task
Describes how to use Cloudera Observability to identify performance trends.

You can identify trends as well as baselines by analyzing your engine's or cluster's performance trends from the Trends chart widget and the Trend tab. Where:

• The Trends time-series chart widget displays more detailed metrics about the processed jobs and queries and enables you to view historical trends for analysis when you select a predefined or custom time period from the Range filter list.
• The Trends tab displays the job or query's instances executed during the selected time range. Depending on the engine, the Trends page displays a job's historical trend from Duration, Data Input, and Data Output histogram charts or lists the runs of the query to show how its performance changes overtime.

Procedure

1. In a supported browser, log in to the Cloudera Observability web UI.
2. In the Clusters page, select the cluster required for analysis.
3. From the Trend chart widget, select the tab of each engine that has processed jobs or queries. Notice if there are any failed or slow jobs or queries.

4. For any Failed Jobs or Queries and for any slow jobs or queries, labeled Job Duration or Query Active Time, select the number under the aforementioned labels.

The following image shows the Total Queries, Failed Queries, and Query Active Time labels from a Workload View's Trend chart widget:
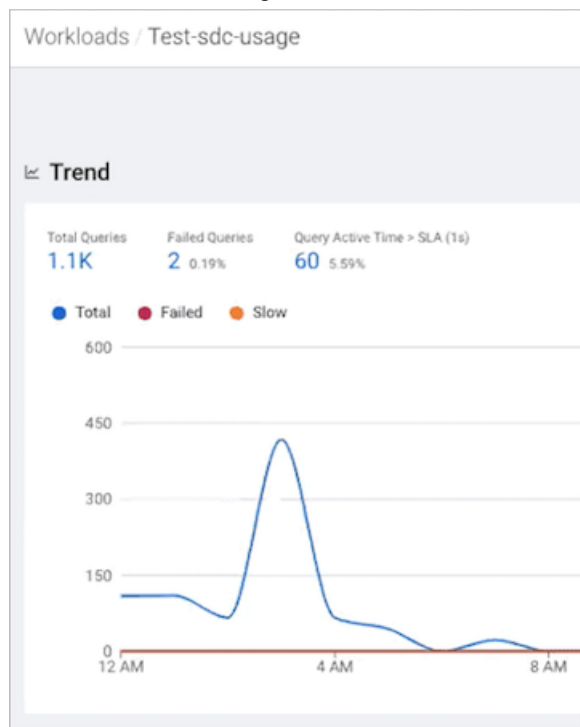
5. From either the Job or Queries page that opens, record those jobs or queries that are failing and their health issue.

6. For more insights, click a job or query and then select the Trend tab.

7. Record your findings.

Use Cloudera Replication Manager to migrate to Cloudera on cloud

Cloudera Replication Manager is a service to copy and migrate data from CDH 5.13+ and above clusters (HDFS, Hive, and HBase data) and Cloudera Base on premises 7.1.4 and above clusters (HDFS , Hive external tables, and HBase data) to Cloudera on cloud clusters. The supported on cloud services include Amazon S3 or Microsoft Azure ADLS Gen2 (ABFS). Replication Manager from HDP clusters to Cloudera on cloud Azure is a beta feature and is not available for general use.

About Replication Manager

Replication Manager is a service in Cloudera on cloud. You can create replication policies in Replication Manager to copy and migrate data from CDH (version 5.13 and higher) clusters (HDFS, Hive, and HBase data) and Cloudera Base on premises (version 7.1.1 and higher) clusters (HDFS, Hive external tables, and HBase data) to Cloudera on cloud clusters. You can also replicate HDFS data from cloud storage to classic clusters (CDH or Cloudera Base on premises clusters), and Hive external tables to Data Hubs. The supported on cloud services include Amazon S3 and Microsoft Azure ADLS Gen2 (ABFS). Replicating Hive managed tables using Replication Manager from HDP clusters to Cloudera on cloud is a beta feature and is not available for general use.

Replication Manager can replicate between services where at least the source service or the destination service is a local service. The term *local service* refers to a service of a cluster managed by Cloudera Manager. In the on-

premises to on-premises replication scenarios, you create the replication policy in the target Cloudera Manager. It is not possible to create replication policies between two non-local services, services on separate peer clusters, or between cloud and a peer cluster's service. In these unsupported scenarios, you can create replication policies on the Cloudera Manager that has the service that you are replicating to/from.

Before you create replication policies, you must ensure that the clusters are supported by Replication Manager. For more information, see Support matrix for Replication Manager on Cloudera on cloud. Ensure that the target cluster's and source cluster's and versions are the same, or the target cluster's version is higher than the source cluster's version.

You can access the Replication Manager service on the Cloudera on cloud web interface. To replicate data between clusters, add the source on-premises clusters as classic clusters on the Management Console Clusters page, add/create one or more Cloudera on cloud SDX Data Lakes and/or Data Hubs, and then create the replication policies in Replication Manager. The **Replication Policies** page shows the progress and status of replication policy jobs. You can also use CDP CLI to create HDFS and Hive replication policies.

Replication Manager provides the following functionalities that you can use to accomplish your data replication goals:

HDFS replication policies

These policies replicate HDFS data and metadata from on-premises clusters (CDH, Cloudera Base on premises, and HDP) to cloud storage buckets such as S3 and ABFS, and from cloud storage to classic clusters (CDH or Cloudera Base on premises clusters). You can choose the frequency of replicating data.

Some use cases where you can use HDFS replication policies include:

- Moving legacy data (from CDH clusters) to cloud deployments (AWS or Azure on Cloudera on cloud).
- Archiving cold data.
- Replicating the required data to another cluster to run analytics on it.

Hive replication policies

These policies support table-level replication and can replicate Hive external tables from on-premises clusters (CDH and Cloudera Base on premises) to cloud storage such as S3 and ABFS and to Data Hubs. They also can:

- replicate data stored in Hive tables, Hive metadata, data in Hive metastore, and Impala metadata (catalog server metadata) associated with Impala tables registered in the Hive metastore, and

    Note: Hive2 managed tables are converted to external tables after replication.

- migrate Sentry permissions to Ranger.

    Note: To perform the Sentry policy replication, you must be running the Sentry service on CDH 5.12 or higher, or any CDH 6.x version.

Some use cases where you can use Hive replication policies include:

- Backing up data periodically.
- Performing a recovery operation when necessary.
- Creating a development and test system for engineers to run quality checks.

HBase replication policies

You can create these policies to replicate HBase data from a source classic cluster (CDH or Cloudera Base on premises cluster), COD, or Data Hub to a target Data Hub or COD cluster. You can also copy or replicate HBase data between different environments within a Virtual Private Cloud (VPC) using these policies. Any future data change in the source cluster is pushed to the target cluster automatically without user intervention.

Some use cases where you can use HBase replication policies include:

- Performing an active-active disaster recovery with conflict resolution (enabling other disaster recovery use cases which provides an efficient utilization of resources).
- Copying required data to the cloud clusters for heavy-duty analytics workloads which helps to optimize on-premises cluster performance.
- Utilizing the continuous data synchronize feature to implement a hybrid cloud that in turn helps you to use it in various other use cases.

CDP CLI for HDFS and Hive replication policies

You can also use CDP CLI commands to create HDFS and Hive replication policies. The CDP CLI commands for Replication Manager are under the replicationmanager CDP CLI option. For more information, see *CDP CLI for Replication Manager*.

CDP CLI is a unified tool to manage all the CDP services. This gives you the flexibility to use it across services from a single pane, view required information in a single scroll (for example, you can view all available clusters' service status in a single page), and collaborate to troubleshoot issues.

Fine-grained permission to access Cloudera Replication Manager

You can restrict access to specific users to view and use Cloudera Replication Manager in a Cloudera on cloud environment so that you can govern the access to critical replication functionalities.

Currently, any user in a Cloudera on cloud environment can view and use Cloudera Replication Manager to create, run, and manage replication policies. However, in some deployments, it is essential that only a few authorized users have access to Replication Manager. This requirement arises when you want to provide an added layer of control which aligns with the best practices for data management and security, and also to enhance security and control over replication management which includes monitoring the replication jobs, and troubleshooting issues efficiently.

Providing role-based access control (RBAC) to Replication Manager users

You can provide fine-grained permissions to specific users to view and use Cloudera Replication Manager.

Procedure

1. Enable the RBAC entitlement. Contact your Cloudera account team to accomplish this task.
2. Identify the users that require access to Cloudera Replication Manager.
3. Go to the Cloudera Management Console User Management page. You can manage the role assignments on this page.

   ⚠️ **Important:** Administrators with the PowerUser role can add, modify, or delete users and groups in the Cloudera on cloud environment.

4. Assign the ReplicationAdmin role to one or more users, and Save the changes.

   🗒️ **Note:** You must have the ReplicationAdmin or PowerUser role to use Replication Manager if the entitlement is enabled.

5. Optionally, create a group to manage the Replication Manager users and their roles. For example, replicationusers.

Accessing Replication Manager UI

You can access the Replication Manager user interface by logging into Cloudera > Select Replication Manager.

When you log into Cloudera Data Platform, the Cloudera on cloud web interface appears. Click Replication Manager. The Overview page of the Replication Manager appears.

The Replication Manager has the following pages:

- Overview

- Classic Clusters
- Cloud Credentials
- Replication Policies

The following image shows the Cloudera on cloud web interface:

You can also access the Replication Manager service by logging into Management Console. In the Management

Console, click and select Replication Manager.

Access Replication Manager in Cloudera on cloud

You can access the Replication Manager service by logging into Cloudera.

When you log into Cloudera, the Cloudera on cloud web interface appears. Click Replication Manager to view the **Overview** page of the Replication Manager.

Replication Manager has the following pages:

- Overview
- Classic Clusters
- Cloud Credentials
- Replication Policies
- Snapshots

The following image shows the Cloudera on cloud web interface:

You can also access the Replication Manager service by logging into Cloudera Management Console. In Cloudera

Management Console, click and select Replication Manager.

Classic Clusters page

The Classic Clusters page specifies the total number of clusters enabled for Replication Manager, the number of clusters that are in an error state, the number of clusters that are active, and the number of clusters for which a warning is issued.

You must register your existing on-premises Cloudera Distribution of Hadoop (CDH) and Cloudera Base on premises clusters on the Cloudera Management Console, after which you can copy or move your data to the cloud. In Cloudera on cloud, these clusters are called *classic clusters*.

The Classic Clusters panel shows the following cluster status:

- **Active** clusters that are currently available to run the replication jobs.
- Clusters having less than 10% disk capacity appears as **Warning**. Click the number to open a table to track the cluster name and the exact disk capacity remaining.
- **Total** number of clusters that are in use.
- Number of clusters that are currently not running as expected or in **Error** state.

To investigate the issues associated with clusters that have an error or warning status, use Cloudera Manager for on-premises clusters.

Cloud Credentials page

The Cloud Credentials page shows the registered cloud credentials for Replication Manager. To replicate data to or from a storage cloud account, you must register the cloud credentials, so that the Replication Manager can access your cloud account. The supported cloud storage accounts are Amazon S3 and Azure Blob Filesystem (ABFS). On the Cloud Credentials page, you can add cloud credentials. You can also update or delete the credentials when necessary.

When you add cloud credentials for your Amazon S3 account, you can choose one of the following authentication methods:

- Access secret key. To use this authentication type, you require an AWS Access Key and an AWS Secret key that you obtain from Amazon. Cloudera Manager stores these values securely and does not store them in world-readable locations. The credentials are masked and encrypted in the configurations passed to processes managed by Cloudera Manager, and redacted from the logs.
- IAM role. Amazon Identity and Access Management (IAM) can be used to create users, groups, and roles for use with Amazon Web Services, such as EC2 and Amazon S3. IAM role-based access provides the same level of access to all clients that use the role.

> **Note:** You can choose the IAM role authentication type only when the following conditions are met:
>
> - The source cluster is hosted on an AWS EC2 infrastructure.
> - The source cluster Cloudera Manager and all the nodes in the cluster are running on an EC2 instance.
> - The source cluster Cloudera Manager has the same IAM role.
>
> For information about configuring AWS credentials, see Introduction to role based provisioning credential in AWS.

You can perform the following tasks on the Cloud Credentials page to manage cloud credentials:

Add cloud credentials

You can add cloud credentials for your S3 or ABFS account. For information about adding cloud credentials, see Working with Cloud Credentials.

> **Note:** Unregistered credentials can impact the replication process. Credentials associated with a cluster node that do not have updated credentials are called unregistered credentials. For example, if a node is down when the credentials are changed on a bucket or when the node is brought up that has the old credentials.

Update cloud credentials

You can update the cloud credentials based on various factors. When the bucket configuration such as secret or access keys, bucket name or endpoint, and encryption type is changed, it can affect the Replication Manager replication policy run and might require an update to the Replication Manager cloud credentials.

Credential changes are picked up by the next run of the policy. When you change the credentials, the in-progress policy runs might fail but the succeeding runs pick up the changes.

To update a cloud credential, click Actions Update .

Delete cloud credentials

You can delete unwanted credentials from the Replication Manager. When you delete cloud credentials, the replication policies that use the deleted cloud credentials might fail. To avoid failures, delete the Replication Manager cloud policies associated with the deleted credentials and recreate the policies with the new credentials. You can view a list of policies associated with specific credentials on the **Cloud Credentials** page.

To delete a cloud credential, click Actions Delete .

Replication Policies page

The "Replication Policies" page shows the number of replication policies that are active, the number of policies that have been suspended, the number of policies that are in error state, and the total number of replication policies available in Replication Manager. The page also provides a detailed view about the replication policies.

The Replication Policies page shows a replication policy status dashboard which shows the following panels and the number of policies with the status:

- **Error** shows the number of replication policies associated with a cluster designated as Error on the Classic Clusters map. Click the number to understand the policy names, the names of the source and destination clusters, and which services are stopped on the source or destination cluster.
- Active replication policies that are in Submitted or Running state. This item is not actionable.
- **Suspended** replication policies that have been suspended by the administrator. This item is not actionable.

- **Total** number of running policies.

Click a replication policy to view more details about the policy. Click Actions ( ) to perform more actions on a replication policy.

Additionally, you can perform the following tasks on the **Replication Policies** page:

- Change the timezone, if required, using .

- View the list of unreachable clusters using .

Replication policy details

You can also view the following policy details on the Replication Policies page:

- Current policy Status .
- Policy Type shows HDFS, Hive, or HBase.
- Replication policy Name .
- Source cluster name.
- Destination cluster name.
- Jobs that were run for the replication policy and its current status.
- Duration or time taken to run the policy.
- Last Success timestamp of the last successful run.
- Next Run timestamp of the next scheduled run.

Optimize Replication Policies page performance

By default, the replication policies are loaded only partially on the **Replication Policies** page, therefore the page might display incomplete statistics about a job status. This is because the job history is necessary to decide whether a policy failed or succeeded. The replication policies with failed jobs might take a longer time to load.

You can change the page load behavior depending on your requirements using . Choose one of the following options to load the **Replication Policies** page faster by delaying to load the job history:

- Delay loading job history when it takes too long attempts to load the job history, but omits the load operation above a certain threshold. By default, Replication Manager uses this option.
- Never load job history minimizes the load on Cloudera Manager and maximizes Replication Manager performance.
- Always load job history ensures that the job history is always loaded for all the displayed replication policies.

Use Case

> Sometimes, Replication Manager fails to reach a healthy Cloudera Manager when there is a temporary networking blip or when there is a load spike on Cloudera Manager. When a cluster becomes unreachable for Replication Manager, the cluster is placed in the list of unreachable
>
> clusters (the list appears when you click). Replication Manager retries to reach the cluster again after 20 minutes. After you confirm that the Cloudera Manager is healthy and expect it to

be reachable by Replication Manager, you can force reload the **Replication Policies** page using

to reconnect every cluster.

For more information, see Replication Policies page does not display all the replication policies.

How replication policies work

In Cloudera Replication Manager, you create replication policies to establish the rules you want applied to your replication jobs. The policy rules you set can include which cluster is the source and which is the destination, what data is replicated, what day and time the replication job occurs, the frequency of job runs, and bandwidth restrictions.

The first time you run a job (an instance of a policy) with data that has not been previously replicated, Replication Manager creates a new folder or database and bootstraps the data. During a bootstrap operation, all data is replicated from the source cluster to the destination. As a result, the initial execution of a job can take a significant amount of time, depending on how much data is being replicated, network bandwidth, and so on. So you should plan the bootstrap operation accordingly.

After the bootstrap operation succeeds, an incremental replication is automatically performed for data replication. This job synchronizes, between the source and destination clusters, any events that occurred during the bootstrap process. After the data is synchronized, the replicated data is ready for use on the destination. Data is in a consistent state only after incremental replication has captured any new changes that occurred during bootstrap.

Subsequent replication jobs from the same source location to the same target on the destination are incremental data replication, so only the changed data is copied.

When a bootstrap operation is interrupted, such as due to a network failure or an unrecoverable error, the Replication Manager does not retry the job instead it runs the job at the next scheduled interval, if available. Therefore, if the bootstrap operation is interrupted, you must manually correct the issue and then run the policy.

When scheduling how often you want a replication job to run, you should consider the recovery point objective (RPO) of the data being replicated; that is, what is the acceptable lag time between the active site and the replicated data on the destination.

Replication policy considerations

You should take into consideration certain guidelines when creating or modifying a replication policy. It is important for you to understand the security restrictions and encryption policies within Replication Manager.

The guidelines you need to consider before you create or modify a replication policy includes:

Data security

> To use an S3 or ABFS cluster for your policy, register your credentials on the Cloud Credentials page. A user with access to the Replication Manager user interface has the ability to browse, within the Replication Manager UI, the folder structure of any clusters enabled for Replication Manager.

> Therefore, users can view folders, files, and databases in the Replication Manager user interface that they might not have access to in HDFS. Users cannot view from the Replication Manager UI the content of files on the source or destination clusters. Nor do these administrators have the ability to modify or delete folders or files that are viewable from the Replication Manager UI.

Policy properties and settings

> Consider the recovery point objective (RPO) of the data being replicated when you schedule a replication policy. The RPO is the acceptable lag time between the active site and the replicated data on the destination. Ensure that the frequency is set so that a job finishes before the next job starts.

> Jobs based on the same policy cannot overlap. If a job is not completed before another job starts, the second job does not execute and is given the status Skipped. If a job is consistently skipped, you might need to modify the frequency of the job.

Specify bandwidth per map, in MBps. Each map is restricted to consume only the specified bandwidth. This is not always exact. The map throttles back its bandwidth consumption during a copy in such a way that the net bandwidth used tends towards the specified value.

Cluster requirements

- Pair the clusters before you include them in a replication policy.
- With a single cluster, you can replicate data on-premises to cloud.
- With a single cluster, you cannot replicate data on-premises to on-premises.
- If the clusters are Replication Manager-enabled, it appears in the Source Cluster or Destination or Data Lake Cluster fields in the Create Policy wizard. You must ensure that the clusters you select are healthy before you start a policy instance (job).

Hive restrictions

- When creating a schedule for a Hive replication policy, you should set the frequency so that changes are replicated often enough to avoid overly large copies.
- ACID tables, managed tables, storage handler-based tables such as Apache HBase, and column statistics are not replicated. ACID tables and managed tables are converted to external tables after replication.

HDFS replication policy

You can use the HDFS replication policies in Cloudera Replication Manager to replicate HDFS data. The HDFS replication policies can replicate HDFS data and metadata from classic clusters (CDH, Cloudera Base on premises, and HDP) to Cloudera on cloud storage buckets such as S3 and ABFS, and from cloud storage to classic clusters (CDH or Cloudera Base on premises clusters). To use an on-premises cluster (CDH or Cloudera Base on premises cluster) in the replication policy, you must register it as a classic cluster in the Management Console. To use the cloud storage for data replication, you must register the cloud credentials in Replication Manager so that the Replication Manager service can access the cloud storage. You must also verify cluster access and configure minimum ports for replication before you create HDFS replication policies.

⚠️ **Important:** Before you create replication policies, see *Support matrix for Replication Manager on Cloudera on cloud* to verify whether your clusters are supported by Replication Manager.

You can also use CDP CLI commands to create HDFS replication policies. The CDP CLI commands for Replication Manager are under the replicationmanager CDP CLI option. For more information, see *CDP CLI for Replication Manager*.

HDFS snapshots

You can schedule taking HDFS snapshots for replication in the Replication Manager. HDFS snapshots are read-only point-in-time copies of the filesystem. You can enable snapshots on the entire filesystem, or on a subtree of the filesystem. In Replication Manager, you take snapshots at a dataset level. Understanding how snapshots work and some of the benefits and costs involved can help you to decide whether or not to enable snapshots.

To improve the performance and consistency of HDFS replications, enable the HDFS replication source directories for snapshots, and for Hive replications, enable the Hive warehouse directory for snapshots. For more information, see HDFS snapshots.

Enabling snapshots on a folder requires HDFS admin permissions because it impacts the NameNode. When you enable snapshots, all the subdirectories are automatically enabled for snapshots as well. So when you create a snapshot copy of a directory, all content in that directory including the subdirectories is included as part of the copy. If a directory contains snapshots but the directory is no longer snapshot-enabled, you must delete the snapshots before you enable the snapshot capability on the directory.

Take snapshots on the highest-level parent directory that is snapshot-enabled. Snapshot operations are not allowed on a directory if one of its parent directories is already snapshot-enabled (snapshottable) or if descendants already contain snapshots.

For example, in the following directory tree image, if directory-1 is snapshot-enabled but you want to replicate subdirectory-2, you cannot select only subdirectory-2 for replication. You must select directory-1 for your replication policy.
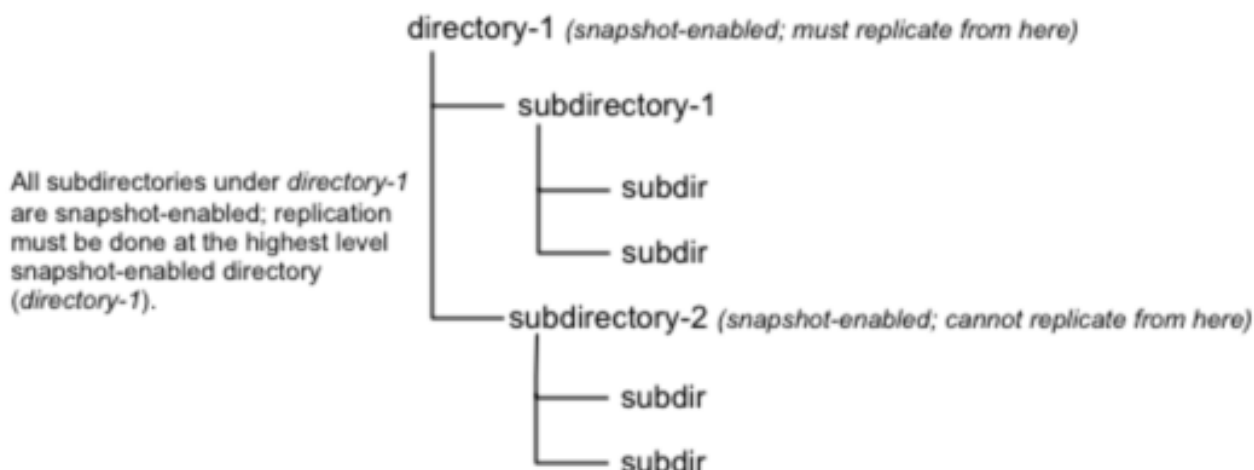
There is no limit to the number of snapshot-enabled directories you can have. A snapshot-enabled directory can accommodate 65,536 simultaneous snapshots. Blocks in datanodes are not copied during snapshot replication. The snapshot files record the block list and the file size. There is no data copying.

When snapshots are initially created, a directory named .snapshot is created on the source and destination clusters under the directory being copied. All snapshots are retained within the .snapshot directories. By default, the last snapshot of a file or directory is retained. Snapshots older than this are automatically deleted. You can configure the number of snapshots to retain when you create or edit an HDFS Snapshot Policy in the target Cloudera Manager using Cloudera Base on premises Replication Manager 7.1.1 or higher or in Backup and Disaster Recovery 5.10 or higher depending on your on-premises target cluster.

Requirements and benefits of HDFS snapshots

You might want to consider the benefits and memory cost of using snapshots. Verify the requirements before you enable snapshots.

Requirements

You must have HDFS superuser privilege to enable or disable snapshot operations. Replication using snapshots requires that the target filesystem data being replicated is identical to the source data for a given snapshot. There must be no modification to the data on the target. Otherwise, the integrity of the snapshot cannot be guaranteed on the target and replication can fail in various ways.

Benefits

Snapshot-based replication helps you to avoid unnecessary copying of renamed files and directories. If a large directory is renamed on the source side, a regular DistCp update operation sees the renamed directory as a new one and copies the entire directory.

Generating copy lists during incremental synchronization is more efficient with snapshots than using a regular DistCp update, which can take a long time to scan the whole directory and detect identical files. And because snapshots are read-only point-in-time copies between the source and destination, modification of source files during replication is not an issue, as it can be using other replication methods.

A snapshot cannot be modified. This protects the data against accidental or intentional modification, which is helpful in governance.

Memory cost

There is a memory cost to enable and maintain snapshots. Tracking the modifications that are made relative to a snapshot increases the memory footprint on the NameNode and can therefore stress NameNode memory. Because of the additional memory requirements, snapshot replication is

recommended for situations where you expect to do a lot of directory renaming, if the directory tree is very large, or if you expect changes to be made to source files while replication jobs run.

Enabling and taking snapshots in Cloudera Manager

Before you take snapshots (in Cloudera Manager) or create snapshot policies (in Replication Manager) for HDFS directories, you must enable snapshots for the directories in Cloudera Manager.

Procedure

1. To enable snapshots for HDFS directories, navigate to the directory on the Cloudera Manager Clusters HDFS service File Browser tab, and click Enable Snapshots.

   Note: If you enable snapshots for a directory, you cannot enable snapshots for its parent directory. Snapshots can be taken only on directories that have snapshots enabled.

   Tip: To disable snapshots for a directory that has snapshots enabled, click Disable Snapshots. Ensure that the snapshots of the directory are deleted before you disable snapshots for the directory.

2. To take a snapshot of a directory or table, perform the following steps:
   a) Navigate to the directory or folder.
   b) Click Take Snapshot in the drop-down menu next to the full file path.
   c) Specify a unique name for the snapshot.

      The snapshot name with the timestamp of its creation appears in the snapshot list.
3. Click Actions Delete to delete a snapshot.

   Note: After you delete a snapshot, you can restore it, if required.

Results

Create a snapshot policy in the Replication Manager to schedule taking a snapshot of the directory at regular intervals.

Hive replication policy

You can create a Hive replication policy in Cloudera Replication Manager after you configure the required Ranger policy in Ranger, register the on-premises cluster (CDH or Cloudera Base on premises) as a classic cluster in Management Console, register cloud account credentials in the Replication Manager service, verify cluster access, and configure minimum ports for replication. The replication load happens on the source on-premises cluster. You can replicate data on-premises to the cloud with a single cluster if the Metastore is running on the cloud.

Replication Manager supports replication of the Hive database from a cluster with underlying HDFS to another cluster with cloud storage. It uses push-based replication, with the replication job running on the cluster with HDFS. For information about Hive cloud replication, see Hive cloud replication

These policies support table-level replication and can replicate Hive external tables from on-premises clusters (CDH and Cloudera Base on premises) to cloud storage such as S3 and ABFS and to Data Hubs. They also can:

• replicate data stored in Hive tables, Hive metadata, data in Hive metastore, and Impala metadata (catalog server metadata) associated with Impala tables registered in the Hive metastore, and

   Note: Hive2 managed tables are converted to external tables after replication.

• migrate Sentry permissions to Ranger.

   Note: To perform the Sentry policy replication, you must be running the Sentry service on CDH 5.12 or higher, or any CDH 6.x version.

Hive metadata replication involves multiple entities. Replication Manager supports replication of external tables in Hive. Hive supports replication of external tables to the target cluster and it retains all the properties of external

tables. The data files permission and ownership are preserved so that the relevant external processes can continue to write in it even after failover.

> ⚠️ **Important:** Before you create replication policies, see Support matrix for Cloudera Replication Manager on page 92 to verify whether your clusters are supported by Replication Manager.

You can also use CDP CLI commands to create Hive replication policies. The CDP CLI commands for Replication Manager are under the replicationmanager CDP CLI option. For more information, see *CDP CLI for Replication Manager*.

The Apache Ranger access policy model consists of the following components:

- Specification of the resources that you can apply to a replication policy which includes the HDFS files and directories; Hive databases, tables, and columns; and HBase tables, column-families, and columns.
- Specification of access conditions for specific users and groups.

You must set the Ranger policy for the hdfs user on the target cluster to perform all operations on all databases and tables. The same user role is used to import Hive Metastore. The hdfs user should have access to all Hive datasets, including all operations. Otherwise, Hive import fails during the replication process.

On the target cluster, the hive user must have Ranger admin privileges. The same hive user performs the metadata import operation.

For more information about Hive replication policies to replicate data from CDH clusters to Cloudera on cloud, see Migrate Hive data from CDH to Cloudera on cloud blog.

Hive replication

Replication Manager allows you to replicate Hive databases from a source cluster to a target location on a destination cluster. The first time you run a job with data that has not been previously replicated, the Replication Manager creates a new folder or database and bootstraps the data. To replicate Hive metadata, Replication Manager performs a full replication. To replicate the data stored in Hive tables, Replication Manager uses snapshot diff-based replication to perform incremental replication.

During bootstrap operation, all of the data from the source location is copied to the destination. This bootstrapping of data can take hours to days, depending on factors such as the amount of data being copied and available network bandwidth.

After the bootstrap operation succeeds, an incremental replication is automatically performed for data replication using snapshot diff-based replication. The job synchronizes, between the source and destination clusters, any events that occurred during the bootstrap process. After the data is synchronized, the replicated data is ready for use on the destination. Data is in a consistent state only after incremental replication has captured any new changes that occurred during bootstrap.

Subsequent replication jobs from the same source location to the same target on the destination are incremental, so only the changed data is copied.

If a bootstrap operation is interrupted, such as due to a network failure or an unrecoverable error, the Replication Manager automatically retries the job. If a retry succeeds, the replication job continues from the point at which it was interrupted. If the automatic retries are not successful, you must manually correct the issue before running the policy again. When you activate the policy again, the replication job resumes from the point at which it was suspended.

Functions such as User Defined Functions (UDF) in Hive are replicated. To enable this, UDFs have to be created using a syntax. An example of UDF creation syntax:

```
CREATE FUNCTION [db_name.]function_name AS class_name  USING JAR|FILE|ARCHIV
E 'file_uri' [, JAR|FILE|ARCHIVE 'file_uri'] ;
```

- ACID tables, external tables, storage handler-based tables (such as HBase), and column statistics are currently not replicated.
- When creating a schedule for a Hive replication policy, you should set the frequency so that changes are replicated often enough to avoid overly large copies.

Snapshot diff-based replication

By default, Replication Manager uses snapshot differences ("diff") to improve performance by comparing HDFS snapshots and only replicating the files that are changed in the source directory.

While Hive metadata requires a full replication, the data stored in Hive tables takes advantage of snapshot diff-based replication. To replicate a database using a Hive replication policy, ensure that all the HDFS paths for the tables in that database are either snapshottable or under a snapshottable root. For example, if the database that is being replicated has external tables, all the external table HDFS data locations should be snapshottable too. Failing to do so can cause the Replication Manager to fail to generate a diff report. Without a diff report, Replication Manager will not use snapshot diff.

An HDFS directory is referred to as snapshottable if an administrator - having superuser privilege or having owner access to the directory - has enabled snapshots for the directory in Cloudera Manager.

You must ensure that the following guidelines are met for efficient incremental replication:

• HDFS snapshots are immutable.

> **Tip:** In the source Cloudera Manager, go to Clusters *HDFS SERVICE* Configuration section, and search for Enable Immutable Snapshots.

• Snapshot root directory is set to as low in the hierarchy as possible.
• Replication Manager user is a super user or the owner of the snapshottable root. This is because the user specified in the Run-as-username field in the replication policy must have the permission to list the snapshots.
• Paths from both source and destination clusters in the replication policy are under a snapshottable root or are snapshottable for the replication policy to run using snapshot diff.

Replication Manager performs a complete replication when one or more of the following change: Delete Policy, Preserve Policy, Target Path, or Exclusion Path.

> **Note:** Ensure the source data does not contain an encrypted subdirectory. This is because snapshot diff-based replication might fail if an encrypted subdirectory exists in the source data.

Hive tables

Managed tables are Hive owned tables where the entire lifecycle of the tables' data are managed and controlled by Hive. External tables are tables where Hive has loose coupling with the data. Replication Manager replicates external tables successfully to a target cluster, and the Hive2 managed tables are converted to external tables after replication.

Hive supports replication of external tables with data to target cluster and it retains all the properties of external tables. The data files' permission and ownership are preserved so that the relevant external processes can continue to write in it even after failover.

The writes on external tables are performed using the Hive SQL commands and the data files can also be accessed and managed by processes outside of Hive. If an external table or partition is dropped, only the metadata associated with the table or partition is deleted but the underlying data files stay intact. A typical example for an external table is to run analytical queries on HBase or Druid owned data using Hive, where the data files are written by HBase or Druid and Hive reads them for analytics.

> **Important:** Hive Materialized Views replication is not supported. However, Replication Manager does not skip it during replication and the replicated data might not work as expected in the target cluster.

When you create a schedule for a Hive replication policy, set the frequency so that changes are replicated often enough to avoid overly large copies.

You might come across the following use cases during Hive replication:

Replication Manager upgrade use case

> In a normal scenario, if you have external tables that are replicated as managed tables, after the upgrade process, you must drop those tables from the target cluster and set the base directory. In the next instance, they get replicated as external tables.

Conflicts in external tables' data location for multiple source clusters replication to the same target cluster

> To handle the conflicts in external tables' data location for multiple source clusters replication to the same target cluster, the Replication Manager assigns an unique base directory for each source cluster under which the external tables' data from the corresponding source cluster is copied.

> For example, if the external table location in a source cluster is /ext/hbase_data, then the location in the target cluster after replication is <base_dir>/ext/hbase_data. You can use the DESCRIBE TABLE command to track the new location of external tables.

Replication conflicts between HDFS and Hive external table location

> When you run the Hive replication policy on an external table, the data is stored on the target directory at a specific location. Next, when you run the HDFS replication policy which tries to copy data at the same external table location, Replication Manager ensures that the Hive data is not overridden by HDFS.

> For example, when you run a Hive replication policy on an external table, the policy creates a target directory /tmp/db1/ext1. When you run an HDFS replication policy, the policy should not override the data by replicating on the /tmp/db1/ext1 directory.

Conflicts during external tables replication process

> Conflicts appear when two Hive replication policies on DB1 and DB2 (either from the same source cluster or different source clusters) have external tables that point to the same data location (for example, /abc) and are replicated to the same target cluster. To avoid such conflicts, you must set different paths for the external table base directory configuration, for both the policies.

> For example, set /db1 for DB1 and /db2 for DB2. This ensures that the target external table data location is different for both databases. For example, /db1/abcd and /db2/abcd.

> **Note:** Replication conflicts are not supported for on-premises to cloud scenario.

Hive cloud replication

Replication Manager supports replication of the Hive database from a cluster with underlying HDFS to another cluster with cloud storage. It uses push-based replication, with the replication job running on the cluster with HDFS.

Hive stores its metadata in Hive Metastore, but the underlying data is stored in HDFS or cloud storage. In a Hadoop cluster with Hive service, the Hive warehouse directory can be configured with either HDFS or cloud storage.

You can perform the following tasks with Hive replication:

• You can rename the dataset in the policy that is replicated.
• You can create a pull-based policy on the source cluster to move data from the target back into the source cluster Hive database.

Hive replication from an HDFS-based cluster to a cloud storage-based cluster requires the following components:

• Source cluster - The cluster with a Hive warehouse directory on local HDFS. This can be an on-premises cluster or an IaaS cluster with data on local HDFS. The required services are HDFS, YARN, Hive, Ranger, Knox, and DLM Engine.
• Destination cluster - The cluster with data on cloud storage. The cluster minimally requires Hive Metastore, Ranger, Knox, and DLM Engine.

Replication Manager does not manage Ranger policies and any Personally Identifiable Information (PII) or any other secure data that gets replicated from on-premises to Amazon S3. You must manage these items outside of Replication Manager.

Table-level replication

To enable table-level replication, you must specify the list of tables to be replicated in a given replication policy. Table-level replication enables you to replicate just the critical tables. It also helps you to speed-up the replication process and also reduces network bandwidth utilization.

You can define table-level replication using regular expressions. You can include or exclude tables in a database for Hive replication during the Hive replication policy creation process.

The following examples illustrate how you can include or exclude Hive tables in the Hive replication policy:

- To include only table1, table2, and table3 in a database for replication, enter the database name in the Database field, and then enter table1|table2|table3 in the Tables field.
- To exclude table5, table7, and table9 and include the rest of the tables in the database, enter the database name in the Database field, and then enter (?!table5|table7|table9).+ in the Tables field.

Limitations using table-level replication

- If a table is dynamically added for replication due to changes in regular expression or added to the include list, the tables' data may not be point-in-time consistent with other tables which are already replicated incrementally. However, this inconsistency is seen for a very small duration until the completion of the next incremental replication after tables are added in the bootstrapped manner.
- Hive does not support overlapping replication policies such as db., db.[t1], and *. to the same target database but works as expected if the target databases are different.

Migrate Sentry authorization policies into Ranger

During Hive replication, Replication Manager migrates Sentry authorization policies into Ranger as part of the replication policy.

The Sentry service serves authorization metadata from the database-backed storage but does not handle actual privilege validation. The Hive and Impala services are clients of this service and it enforces Sentry privileges when the services are configured to use Sentry. Replication Manager allows administrators to migrate the existing Sentry permissions from the source CDH cluster to the Ranger policies in Cloudera on cloud.

When you create a replication policy, you can choose to migrate the Sentry policies for the resources that you want to migrate. During replication policy job run, the resources and its Sentry policies are migrated to the destination cluster. To migrate the Sentry policies for the resources, select the Include Sentry Permissions with Metadata option in the Additional Settings page of the Create Replication Policy wizard.

Note: To perform the Sentry policy replication, you must be running the Sentry service on CDH 5.12 or higher, or any CDH 6.x version.

The Sentry Permissions section of the Create Replication Policy wizard has the following options:

- Include Sentry Permissions with Metadata migrates the Sentry permissions during the replication job run.
- Exclude Sentry Permissions with Metadata ensures that the Sentry permissions are not migrated during the replication job.
- Choose Skip URI Privileges if you do not want to include URI privileges when you migrate Sentry permissions. During migration, the URI privileges are translated to point to an equivalent location in S3. If the resources have a different location in Amazon S3, do not migrate the URI privileges because the URI privileges might not be valid.

The following image shows the Sentry Permissions section in the Create Replication Policy wizard:

The following steps are completed during the migration of Sentry policies into Ranger:

- The Export operation runs in the source cluster. During this operation, the Sentry permissions are fetched and exported to a JSON file. This file might be in a local file system or HDFS or S3, based on the configuration.
- The Translate and Ingest operations take place on the target cluster. In the translate operation, Sentry permissions are translated into a format that can be read by Ranger. The permissions are then imported into Ranger. When the permissions are imported, they are tagged with the source cluster name and the time that the ingest took place. After the import, the file containing the permissions is deleted.

  Note: During Hive replication from an on-premises CDH cluster to a cloud cluster, the Replication Manager migrates Sentry authorization policies into Ranger as part of the replication policy. However, no import operation is initiated if the end service in the cloud cluster (AWS) is Sentry.

A Ranger policy is created for each resource, such as a database, table, or column. The policy name is derived from the resource name. For example, if the resource is Database:dinosaurs, table= theropods, then the derived policy name is database=dinosarus->table=theropods.

The priority for migrated policies is set to normal in Ranger. The normal priority allows you to create another policy for the same resource that overrides the policy that is imported from Sentry.

Sentry to Ranger permissions
There are no one-to-one mapping between Sentry privileges and Ranger service policies, therefore the Sentry privileges are translated to their equivalents within the Ranger service policies.

The following list illustrates how the Sentry privileges appear in Ranger after the migration:

- Sentry permissions that are granted to roles are granted to groups in Ranger.
- Sentry permissions that are granted to a parent object are granted to the child object as well. The migration process preserves the permissions that are applied to child objects. For example, a permission that is applied at the database level also applies to the tables within that database.
- Sentry OWNER privileges are translated to the Ranger OWNDER privilege.

- Sentry OWNER WITH GRANT OPTION privileges are translated to Ranger OWNER with Delegated Admin checked.
- Sentry does not differentiate between tables and views. When view permissions are migrated, they are treated as table names.
- Sentry privileges on URIs uses the object store location as the base location.
- If your cluster contains the Kafka service and the Kafka sentry policy had "action": "ALL" permission, the migrated Ranger policy for "cluster" resource will be missing the "alter" permission. This is only applicable for "cluster" resource. You will need to add the policy manually after the upgrade. This missing permission will not have any functional impact. Adding the "alter" permission post upgrade is needed only for completeness because the 'configure' permission will allow alter operations.

The table below shows how actions in Sentry are applied to the corresponding action in Ranger:

Table 1: Sentry Actions to Ranger Actions

Sentry Action	Ranger Action
SELECT	SELECT
INSERT	UPDATE
CREATE	CREATE
REFRESH	REFRESH
ALL	ALL
SELECT with Grant	INSERT
INSERT with Grant	INSERT
CREATE with Grant	CREATE
ALL with Grant	ALL with Delegated Admin Checked

HBase replication policy

You can replicate HBase and Phoenix tables using HBase replication policies in Cloudera Replication Manager. An HBase replication policy replicates the data at table-level granularity. After you create an HBase replication policy, you can delete one or more tables from the policy. The replication policy replicates the data in the specified tables and continues to replicate the generated data (that is, future changes in data) unless you suspend the policy or delete the tables.

The replication policy replicates the data in the specified tables and continues to replicate the generated data (that is, future changes in data) unless you suspend the policy or delete the tables. You can replicate only existing HBase data, generated HBase data, or both depending on your requirements. You also can choose to replicate all the HBase tables or only the required tables in a database.

Before you create an HBase replication policy, you must:

- verify whether your clusters are supported by Replication Manager.
- understand how first-time setup configuration works.
- understand how cluster pairing works.
- understand the available methods to replicate HBase data.

You can also replicate HBase data simultaneously between multiple clusters.

Supported clusters for HBase replication policies
Before you create HBase replication policies, you must verify whether your clusters are supported by Replication Manager.

The following image shows the lowest supported versions that you can use for HBase replication policies. Consult the Support matrix for Replication Manager on Cloudera on cloud for the complete list of supporting clusters and scenarios:

Figure 1: Lowest supported versions for HBase replication policies
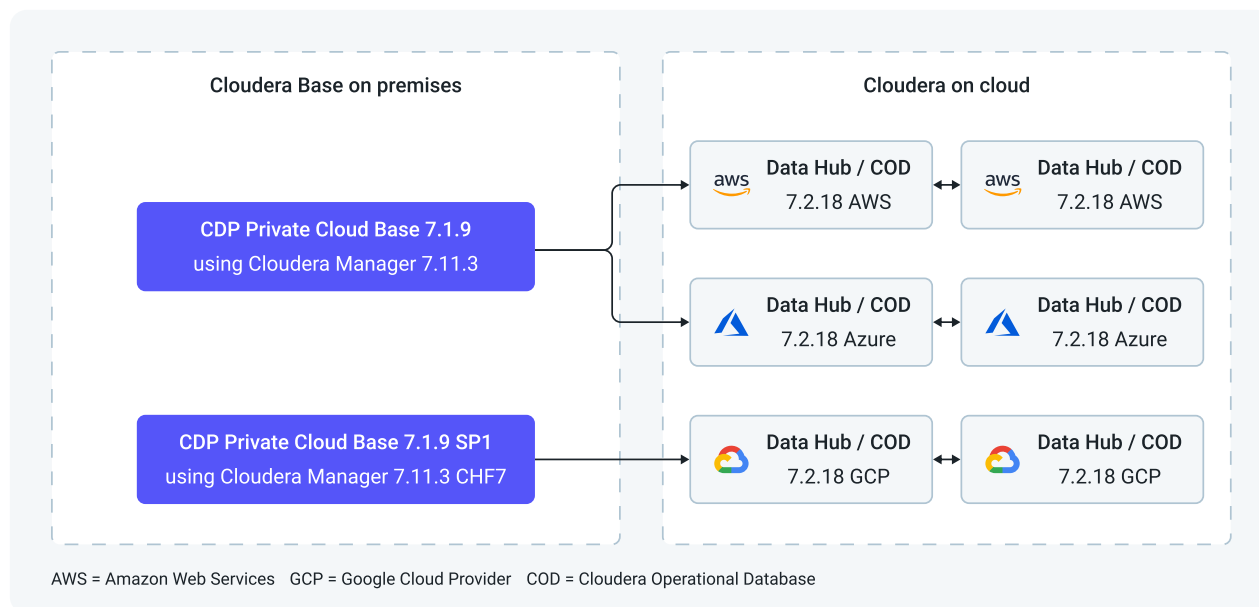
How HBase replication policies work

After you create the first HBase replication policy between a source cluster and target cluster, the Replication Manager service starts several background processes. After the processes complete, the service initiates the HBase data replication. One of the main background process is the first-time setup configuration.

- Understanding first-time setup configuration process
- What is a cluster pair

Steps in first-time setup configuration process

When you create the first HBase replication policy to replicate HBase data from a source cluster to a target cluster, the Replication Manager completes the first-time setup configuration steps and then replicates the data. The first-time setup configuration between a cluster pair is a one-time process, therefore subsequent HBase replication policies for the same cluster pair (ClusterA and ClusterB) do not require a first-time setup.

The first-time setup configuration completes the following steps:

1. Creates peers between the source (ClusterA) and target (ClusterB), that is, creates a cluster pair between ClusterA and ClusterB.
2. Initiates the required configuration changes. One of the steps ensures that both the clusters use the same credentials.jceks file.
3. Restarts the HBase service on both the clusters.

> ⚠️ **Important:** If you are replicating HBase data from on-premises cluster to Cloudera Operational Database or Cloudera Data Hub, you must manually restart HBase service on the source cluster.

When HBase replication policies are created, modified, or deleted simultaneously, Replication Manager processes each operation independent of each other.

> 🖊️ **Note:** When the first-time setup between two clusters is in progress, you can create HBase replication policies between them. However, you cannot use one of these clusters with another cluster to create an HBase replication policy (in multi-cluster replication scenario) until the first-time setup is complete.

What is a cluster pair

The first-time setup configuration creates peers between the source (ClusterA) and target (ClusterB), that is, creates a cluster pair between ClusterA and ClusterB. If a cluster pair (ClusterA and ClusterB) has one or more active/ suspended HBase replication policies between them, you cannot pair either of the clusters with another cluster.

You can use ClusterA or ClusterB with another cluster in an HBase replication policy only if the following conditions are true:

- ClusterA and ClusterB do not have existing HBase replication policies.
- All existing active/suspended HBase replication policies are deleted.

A warning appears on the **Select Destination** page during HBase replication policy creation if you choose ClusterA or ClusterB as source or target (when no HBase replication policies exist between them) with another cluster or if the other cluster in the cluster pair is unreachable. To continue HBase replication policy creation, you must acknowledge the forced re-pairing of the clusters.

Methods to replicate HBase data

You can replicate only the existing HBase data, generated HBase data, or both depending on your requirements. You also can choose to replicate all the HBase tables or only the required tables in a database.

When you create an HBase replication policy, you can choose one or more of the following methods to replicate HBase data depending on your requirements:

- Replicate only the generated data from chosen tables.
- Replicate existing data and generated data from chosen tables.
- Replicate existing tables and future tables in the database.
- Replicate existing data and generated data from chosen tables and future tables

Replicate only the generated data from chosen tables

In this method, you choose one or more tables during the replication policy creation process. Replication Manager replicates only the data that is generated after policy creation.

Replicate existing data and generated data from chosen tables

In this method, you choose one or more tables, and also choose the Select Source Perform Initial Snapshot option during the HBase replication policy creation process. Replication Manager replicates the existing data and the data that is generated after policy creation.

For example, you have two tables named 'Orders' and 'Customers' in the source cluster and you want to copy the data from these tables from March 1, 2021 onwards. To accomplish this task, you create an HBase replication policy without choosing the Perform Initial Snapshot option in the **Create Replication Policy** wizard on March 1, 2021. The data that you create, update, or delete in the source cluster after you created the HBase replication policy is automatically replicated to the target cluster.

Replicate existing tables and future tables in the database

In this method, you choose the Select Source Replicate Database option during the HBase replication policy creation process. Replication Manager replicates the generated data from the existing tables, and it replicates data from the future tables that are created after the HBase replication policy creation process is complete.

To replicate data from the future tables successfully, you must create similar empty tables on the target cluster. You can perform this action when you create or add a table to the database on the source cluster.

You can choose the Replicate Database option only if the following conditions are true:

- Target Cloudera Manager version is 7.11.0 or higher.
- Source cluster's CDH version is 6.x or higher.

  CDH 5.16.2 and higher versions also support the Replicate Database option after you upgrade the source cluster Cloudera Manager.

- No existing HBase replication policies exist between the source and target clusters.

> **Tip:** If you want to replicate the new tables that are created after the replication policy creation is complete, you must configure the replication scope to "1" for those tables on the source cluster.
>
> To configure the replication scope for a table on the master cluster, run the alter *[***TABLE NAME***]*, {NAME =>   *[***COLUMN FAMILY***]*, REPLICATION_SCOPE   => 1} command for each column family that must be replicated. *REPLICATION_SCOPE* is a column-family level attribute, where the value '0' means replication is disabled, and '1' means replication is enabled.

After you select the Select Source Replicate Database option in the HBase replication policy wizard, you can choose one of the following options to determine the tables in the database to replicate:

- Replicate all user tables - Replicates all the HBase tables in the database after the replication scope of the tables are set to 1.
- Replicate only tables where replication is already enabled - Replicates only those tables for which the replication scope is already set to 1.

  This option is supported only if the target cluster Cloudera version is 7.2.17.300 using Cloudera Manager 7.11.0-h3 or higher versions or Cloudera version 7.2.16.500 using Cloudera Manager 7.9.0-h7 or higher versions, or Cloudera version 7.12.0.0.

Replicate existing data and generated data from chosen tables and future tables

In this method, you choose the Perform Initial Snapshot and Replicate Database options on the **Select Source** page during the HBase replication policy creation process. You can also choose to replicate all the tables in the database or only those tables for which the replication scope is already set to 1. Replication Manager replicates the existing and generated data from the existing tables in addition to the data in future tables.

Replicate HBase data simultaneously between multiple clusters
Starting from Cloudera on cloud 7.2.16.500, 7.2.17.200, and 7.2.18, you can create multiple HBase replication policies between multiple clusters to replicate HBase data. You must consider the limitations before you create a multi-cluster replication scenario. You can use the multi-cluster replication scenario for various use cases.

- How multi-cluster HBase replication works
- Limitations
- Sample use cases

How multi-cluster HBase replication works

The first-time setup configuration consists of several steps of which one step is to ensure that the source and target cluster use the same credentials.jceks file. Therefore, if multiple supported clusters share the same credentials.jceks file, you can replicate HBase data between them seamlessly using HBase replication policies.

The following image shows a sample multi-cluster HBase data replication scenario and a few possible directions of replication:
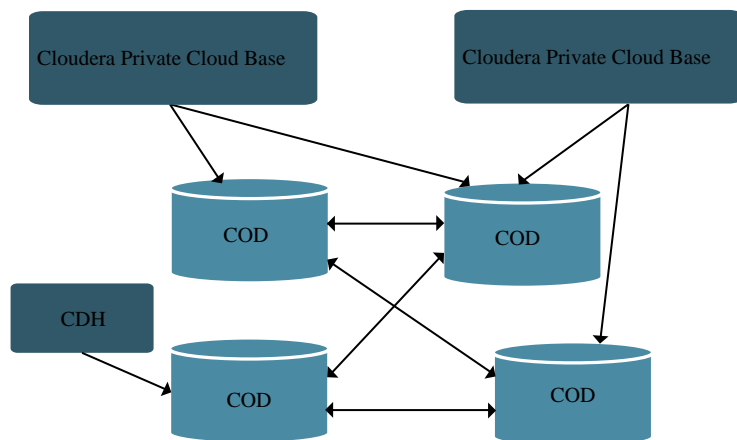
It is recommended that you do not replace the credentials.jceks file manually to create a multi-cluster HBase replication scenario. This is because when you create the first HBase replication policy between a pair of clusters, Replication Manager triggers the first-time setup process during which the credentials.jceks file in both the clusters get synchronized as required for HBase data replication.

Limitations

Consider the following limitations before you replicate the HBase data between multiple clusters using HBase replication policies:

- An HBase replication policy in a multi-cluster HBase replication setup fails when you use clusters that are part of another independent replication setup. This is because the clusters use a different credentials.jceks file. To use these clusters, you must break the cluster pairing and then create the required HBase replication policies.

  > **Tip:** To break the cluster pairing, use the `POST /dmx/api/clusters/<target cluster crn>/hbase/resetFirstTimeSetup` API with the `{"sourceCluster": "<source cluster data center>$<source cluster name>"}` payload.

  Monitor the growing multi-cluster replication network so that it does not get disconnected. This ensures that the credentials.jceks file is the same on all clusters, the replication setup is always consistent, and no existing replication scenarios have to be reset.

- The Replication Manager UI does not allow the HBase replication policy creation to proceed if you choose a cluster (as source or target) that is in another first-time setup process. In this instance, you can wait for a few minutes to allow the first-time setup to complete and then create the HBase replication policy.

  When you create the first HBase replication policy between two clusters, the first-time setup configuration is initiated. After the configuration completes, the HBase data replication is initiated.

  > **Important:** In a multi-cluster replication scenario, when you use a source on-premises cluster, you must manually restart the HBase service on the cluster after the first-time setup configuration completes. However, the next time you choose the same source cluster with another target cluster, the manual restart is not required. The same is true for the automatic HBase service restart in Data Hubs where the restart is performed only when the first replication is created with a particular cluster.

- The following conditions must be met to use the IDBroker credentials to create multiple HBase replication policies between multiple clusters when the target COD clusters are in separate AWS accounts or when a single AWS Role does not have access to all the required S3 buckets for all HBase target clusters:

  - Use Cloudera on cloud 7.2.18.200 or higher versions.
  - Choose the Perform Initial Snapshot option, and then specify the custom username in the Export snapshot user field in the **Select Source** page during the HBase replication policy creation process.

Use cases

Some use cases where you can use the multi-cluster HBase replication scenarios are illustrated below:

- Multiple source clusters and a single target cluster. You might have a disaster-recovery use case where you want to use a single COD to back up all the HBase data. The following image illustrates this scenario:
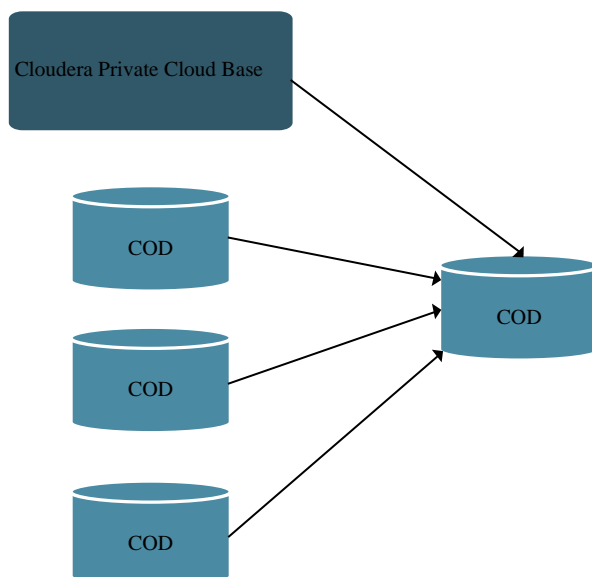
- Single source cluster and multiple target clusters. You might have a use case where all the HBase data is located in a cluster and you want to replicate only specific HBase tables to different environments to fulfill specific requirements. For example, QE environments and/or experimentation use case. The following image illustrates this scenario:
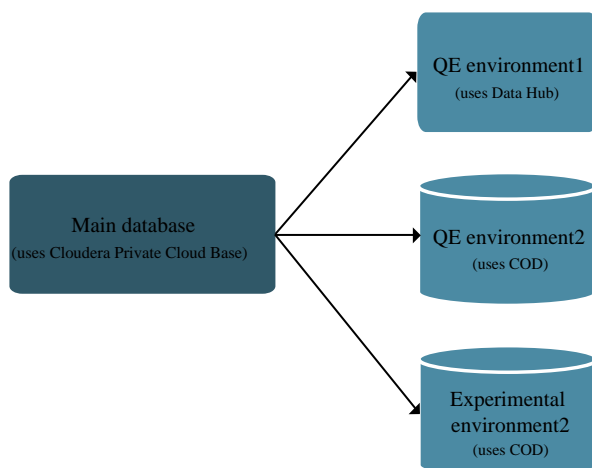
Using HDFS replication policies

You can use the HDFS replication policies in Cloudera Replication Manager to replicate HDFS data. The HDFS replication policies can replicate HDFS data and metadata from classic clusters (CDH, Cloudera Base on premises, and HDP) to Cloudera on cloud storage buckets such as S3 and ABFS, and from cloud storage to classic clusters (CDH or Cloudera Base on premises clusters). To use an on-premises cluster (CDH or Cloudera Base on premises) in

the replication policy, you must register it as a classic cluster in the Cloudera Management Console. To use the cloud storage for data replication, you must register the cloud credentials in Replication Manager so that the Replication Manager service can access the cloud storage. You must also verify cluster access and configure minimum ports for replication before you create HDFS replication policies.

> ⚠️ **Important:** Before you create replication policies, see *Support matrix for Replication Manager on Cloudera on cloud* to verify whether your clusters are supported by Replication Manager.

You can also use Cloudera CLI commands to create HDFS replication policies. The Cloudera CLI commands for Replication Manager are under the replicationmanager Cloudera CLI option. For more information, see *Cloudera CLI for Replication Manager*.

Preparing to create an HDFS replication policy

Before you create the HDFS replication policies in Cloudera Replication Manager to replicate HDFS data, register the on-premises cluster (CDH or Cloudera Base on premises) as a classic cluster in Management Console, register cloud account credentials in the Replication Manager service, verify cluster access, and configure minimum ports for replication.

Procedure

- Do the source cluster and target cluster meet the requirements to create an HDFS replication policy?

  > 🗒️ **Note:** Ensure that the source and target clusters' hostnames are different.

  The following image shows the support matrix based on the lowest supported End of Support (EoS) versions that you can use for HDFS replication policies. Consult the Support matrix for Cloudera Replication Manager for the complete list of supporting clusters and scenarios:

  Figure 2: Lowest supported versions for HDFS replication policies
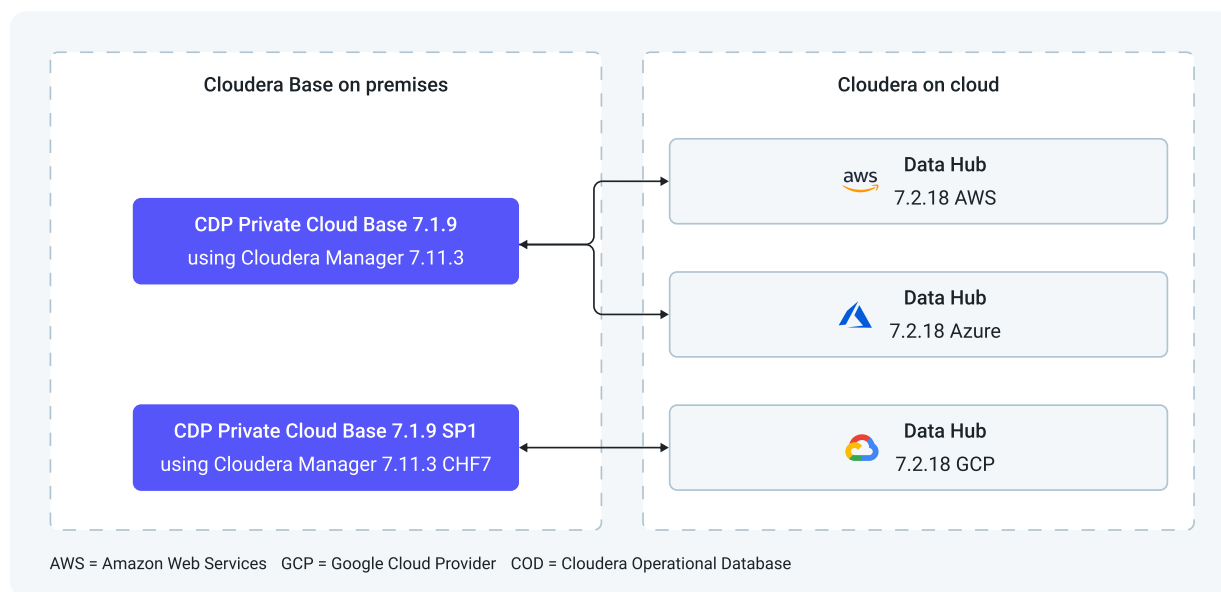
- Is the required on-premises cluster (CDH cluster or Cloudera Base on premises cluster) registered as a classic cluster on the Management Console?

  CDH clusters and Cloudera Base on premises clusters are managed by Cloudera Manager. To enable these on-premises clusters for Replication Manager, you must register them as Classic Clusters on the Management Console. After registration, you can use them for data migration purposes.

  > ⚠️ **Important:** When you register a Cloudera Base on premises cluster as a classic cluster, ensure that you use only the Cloudera Manager IP Address and Cloudera Manager Port options and not the Register KNOX endpoint (Optional) option.

  For information about registering an on-premises cluster as a classic cluster, see Adding a CDH cluster and Adding a Cloudera Base on premises cluster.

- Is an external account available in the Cloudera Manager instance that has access to the bucket or container that you are using in the HDFS replication policy?

  For more information, see Role-based credential on AWS, App-based credential on Azure, and Cloudera Manager documentation.

- Do you have the required cluster access to create replication policies?

  Power users, the user who onboarded the source and target clusters, and users with ClassicClusterAdmin or ClassicClusterUser resource roles can create replication policies on clusters for which they have access. For more information, see Understanding account roles and resource roles.

  > 💡 **Tip:** Ensure that you have *Replication Administrator* or Full Administrator role on the source on-premises cluster.

- Do you have the required cluster access to view the replication policies?

  Existing HDFS replication policies are visible to users who have access to the source cluster in the replication policy. A warning appears if you do not have access to the source cluster.

  If you can view the policies, you can perform other actions on the policy including policy update and policy delete operations.

  > 📝 **Note:** A Replication Manager user can browse, within Replication Manager, the folder structure of any cluster enabled for Replication Manager. Therefore, permitted users can view folders, files, and databases in Replication Manager that they might not have access to in HDFS. The users cannot view the content of files on the source or destination clusters, nor can they modify or delete folders or files that are viewable from Replication Manager.

- Is the required cloud credential that you want to use in the replication policy registered with the Replication Manager service?

  For more information, see Working with cloud credentials.

- Are the following ports open and available for Replication Manager?

Table 2: Minimum ports required for HDFS replication policies

Connectivity required	Default Port	Type	Description
Data transfer from classic cluster hosts to cloud storage	80 or 443 (TLS)	Outbound	Outgoing port. All classic cluster nodes must be able to access S3/ ADLS Gen2 endpoint.

Connectivity required	Default Port	Type	Description
Classic cluster	6000-6049 for CCMv1 443 for CCMv2	Outbound	Connecting source classic cluster to the Cloudera Management Console through Cluster Connectivity Manager (CCM). For more information, see Outbound network access for CCM, and CCM overview.

Consider the following best practices while using on Microsoft Azure ADLS Gen2 (ABFS):

- Ensure that the on-premises cluster (port 443) can access the *https://login.microsoftonline.com* endpoint. This is because the Hadoop client in the on-premises cluster (CDH/) connects to the endpoint to acquire the access tokens before it connects to Azure ADLS storage. For more information, see the *General Azure guidelines* row in the Azure-specific endpoints table.
- Ensure that the steps mentioned in the *General Azure guidelines* and *Azure Data Lake Storage Gen 2* rows in the Azure-specific endpoints table are complete so that the endpoint connects to the target path successfully.

The following system architecture diagram shows the interaction between components during HDFS replication using HDFS replication policies:

Figure 3: System architecture diagram for HDFS replication in Cloudera Replication Manager
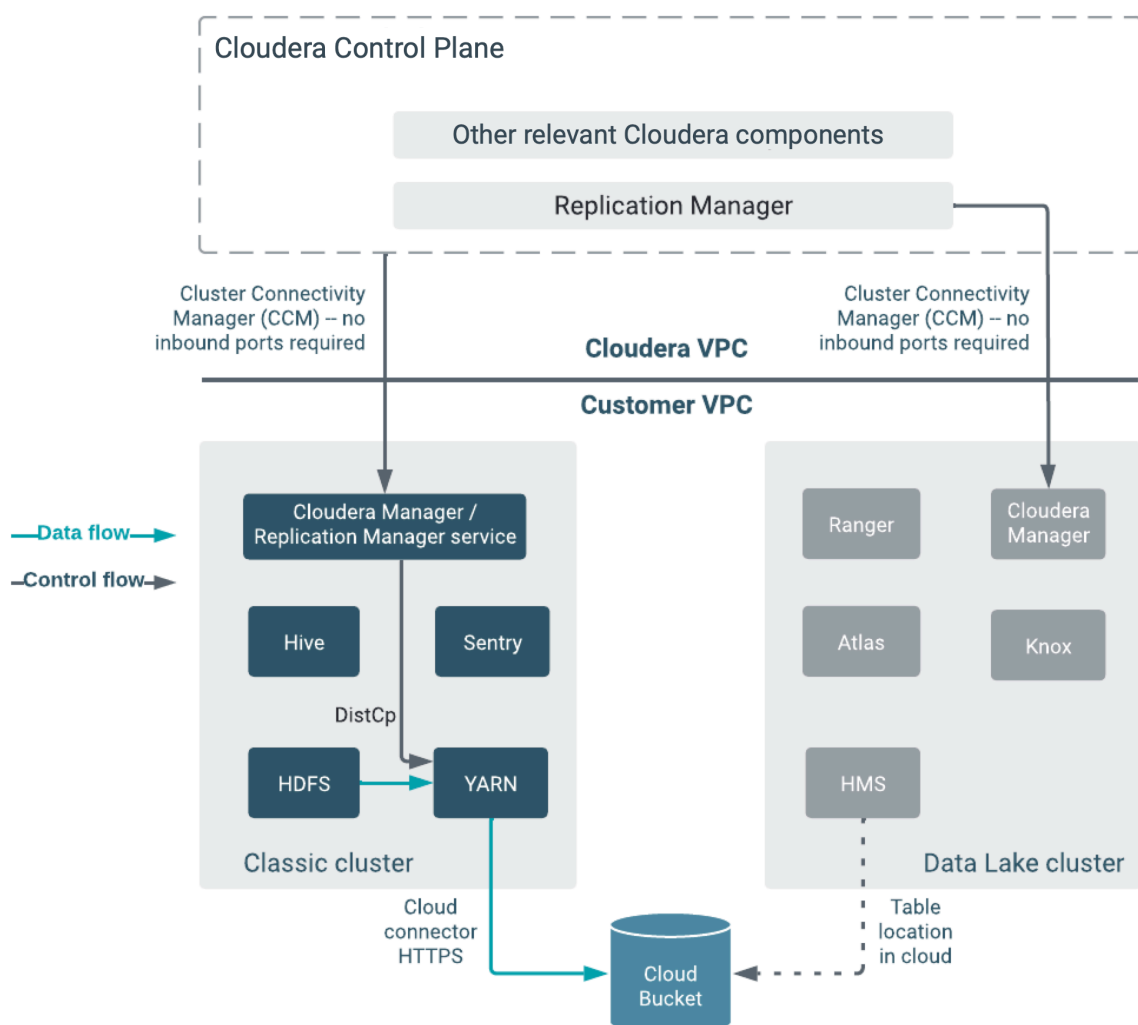
What to do next

After the clusters and cloud storage requirements are met, you can create an HDFS replication policy.

Creating HDFS replication policy

After you register the classic clusters in Management Console and register the cloud credentials in Cloudera Replication Manager, you can create an HDFS replication policy to replicate HDFS data. An HDFS replication policy can replicate HDFS data from a classic cluster (CDH, Cloudera Base on premises, and HDP clusters) to cloud storage, and from cloud storage to classic clusters (CDH or Cloudera Base on premises clusters). You can replicate data on-premises to cloud storage account with a single cluster.

About this task

Alternatively, you can also use CDP CLI commands to create HDFS replication policies. The CDP CLI commands for Replication Manager are under the replicationmanager CDP CLI option. For more information, see CDP CLI for Replication Manager.

Procedure

1. On the Cloudera Management Console Replication Manager Replication Policies page, click Add Policy.

   The **Create Replication Policy** wizard appears.

2. On the General page, choose or enter the following information:

Option	Description
HDFS	Creates an HDFS replication policy.
Policy Name	Enter a unique name for the replication policy.
Description	Optional. Enter a brief description about the replication policy.

   The following image shows a sample General page in the **Create Replication Policy** wizard:

3. Click Next.

4. On the Select Source page, the options change depending on whether you choose a classic cluster or cloud storage as the source for data to be replicated.

   a) If you are replicating from a classic cluster , choose or enter the following information:

Option	Description
Type	Choose CLUSTER to select an on-premises cluster as the source cluster.
Source Cluster	Select a classic cluster.
Source path	Choose one of the following methods to determine the directory where source data resides on the source cluster: • Enter the complete directory path. • Click File Browser to view and navigate the existing directory list on the selected cluster. Select the required directory that you want to replicate.
Run As Username (on source)	Optional. The replication policy uses the Default username to replicate HDFS data. If you are using a kerberized cluster, enter the required username. The replication policy uses this username to replicate the data in the kerberized cluster.

   b) If the source of your replication is cloud storage, choose or enter the following information:

Option	Description
Type	Choose S3 or ABFS to select your cloud provider.
Cloud Credential on Source	Choose the required cloud credential. The cloud credentials that you register for Replication Manager on the Cloud Credentials page appear in this field.
Path	Based on the cloud provider, enter the path to the source data in the required format. • For an S3 bucket, provide the path of the directory in the *[***BUCKET NAME***]/[***PATH***]* format. • For an ABFS container, provide the path of the directory in the *[***FILESYSTEM***]@[***STORAGE ACCOUNT***]/[***LOCATION***]* format.

5. Click Next.

6. On the Select Destination page, the options change depending on whether you choose a classic cluster or cloud storage as the destination for the replicated data:

   a) If you are replicating from a classic cluster to cloud storage, choose or enter the following information:

Option	Description
Type	Choose S3 or ABFS to select your cloud provider.
Cloud Credential on Source	Choose the required cloud credential. The cloud credentials that you register for the selected source cluster on the **Cloud Credentials** page appear in this field. You can also add cloud credentials using the Add Cloud Credential link.
Path	Based on the cloud provider, enter the target path, where data is replicated to, in the required format. • For an S3 bucket, provide the path of the directory in the *[***BUCKET NAME***]/[***PATH***]* format. • For an ABFS container, provide the path of the directory in the *[***FILESYSTEM***]@[***STORAGE ACCOUNT***]/[***LOCATION***]* format.

b) If you are replicating from cloud storage to a classic cluster, choose or enter the following information:

Option	Description
Type	Choose Classic Cluster.
Destination Cluster	Choose a classic cluster as the target cluster.
Destination Path	Enter the directory on the target cluster to which the replication policy replicates HDFS data.
Run As Username	Optional. The replication policy uses the Default username to replicate HDFS data. Enter another username if you want the replication policy to use it to replicate data.

7. Click Validate Policy.

Replication Manager verifies whether the details provided are correct.

8. Click Next.

9. On the **Schedule** page, choose or enter the following information:

Option	Description
Run Now	Starts to replicate the existing HDFS data after the replication policy creation is complete. Choose the frequency to replicate data periodically.
Schedule Run	Runs the replication policy to replicate data at a later time. Choose the date and time for the first run, and then choose the frequency to replicate data periodically. **Tip:** On the **Replication Policies** page, click to change the timezone.
Frequency	Choose one of the following options: • Does Not Repeat • Custom - In the Custom Recurrence dialog box, choose the time, date, and the frequency to run the policy. Replication Manager ensures that the same number of seconds elapse between the runs. For example, if you choose the Start Time as January 19, 2022 11.06 AM and Interval as 1 day, Replication Manager runs the replication policy for the first time at the specified time in the timezone the replication policy was created in, and then runs it exactly after 1 day that is, after 24 hours or 86400 seconds. **Note:** Ensure that the frequency in a schedule enables a job to finish before the next job starts. Also, ensure that the jobs based on the same policy do not overlap. If a job is not completed before another job starts, the second job does not run and the job status appears as Skipped. If a job is consistently skipped, you might need to modify the frequency of the job.

10. Click Next.

11. On the Additional Settings page, enter or choose the values as necessary:

Option	Description
YARN Queue Name	Enter the name of the YARN queue for the cluster to which the replication job is submitted if you are using Capacity Scheduler queues to limit resource consumption. The default value for this field is default.

Option	Description
Maximum Maps Slots	Set the maximum number of map tasks (simultaneous copies) per replication job. The default value is 20.
Maximum Bandwidth	Adjust this setting so that each map task is throttled to consume only the specified bandwidth. Each map task ((simultaneous copy) is restricted to consume only the specified bandwidth. This is not always exact. The map throttles back its bandwidth consumption during a copy in such a way that the net bandwidth used tends towards the specified value. You can adjust this setting so that each map task is throttled to consume only the specified bandwidth so that the net bandwidth used tends towards the specified value. The default value for the bandwidth is 100MB per second for each mapper.
Path Exclusion	Enter one or more regular expressions separated by comma. Replication Manager does not copy the subdirectories or files from the source that matches one of the specified regular expressions to the target cluster.
Replication Strategy	Choose one of the following replication strategies to determine whether the file replication tasks should be distributed among the mappers statically or dynamically. • Static distributes file replication tasks among the mappers up front to achieve an uniform distribution based on the file sizes. • Dynamic distributes the file replication tasks in small sets to the mappers, and as each mapper completes its tasks, it dynamically acquires and processes the next unallocated set of tasks. The default replication strategy is Dynamic.
MapReduce Service	Choose the MapReduce or YARN service to use. **Note:** When the target cluster of the replication policy is part of a multi-cluster Cloudera Manager, you must ensure that you select the configured YARN service from the same cluster where the replication's target service is deployed.
Log Path	Enter an alternate path for the logs, if required.

Option	Description

Error Handling	Select the following options as necessary:
	
	 • Skip Checksum Checks - Determines whether to skip checksum checks on the copied files. If selected, checksums are not validated. Checksums are checked by default.
	
	     Note: You must skip checksum checks to prevent replication failure due to non-matching checksums in the following cases:
	
	       • Replications from an encrypted zone on the source cluster to an encrypted zone on a destination cluster.
	      • Replications from an encryption zone on the source cluster to an unencrypted zone on the destination cluster.
	      • Replications from an unencrypted zone on the source cluster to an encrypted zone on the destination cluster.
	
	   Checksums are used for two purposes:
	
	     • To skip replication of files that have already been copied. If Skip Checksum Checks is selected, the replication job skips copying a file if the file lengths and modification times are identical between the source and destination clusters. Otherwise, the job copies the file from the source to the destination.
	    • To redundantly verify the integrity of data. However, checksums are not required to guarantee accurate transfers between clusters. HDFS data transfers are protected by checksums during transfer and storage hardware also uses checksums to ensure that data is accurately stored. These two mechanisms work together to validate the integrity of the copied data.
	
	 • Skip Listing Checksum Checks - Whether to skip checksum check when comparing two files to determine whether they are same or not. If skipped, the file size and last modified time are used to determine if files are the same or not. Skipping the check improves performance during the mapper phase. Note that if you select the Skip Checksum Checks option, this check is also skipped.
	
	 • Abort on Error - Whether to abort the job on an error. If selected, files copied up to that point remain on the destination, but no additional files are copied. Abort on Error is not selected by default.
	
	 • Abort on Snapshot Diff Failures - If a snapshot diff fails during replication, the replication policy uses a complete copy to replicate data. If you select this option, the policy aborts the replication when it encounters an error instead.

Option	Description
Preserve	Choose the required options to preserve the block size, replication count, permissions (including ACLs), and extended attributes (XAttrs) as they exist on the source file system, or to use the settings as configured on the destination file system. By default source system settings are preserved. When Permission is selected, and both the source and destination clusters support ACLs, replication preserves ACLs. Otherwise, ACLs are not replicated. When Extended attributes is selected, and both the source and destination clusters support extended attributes, replication preserves them. (This option only displays when both source and destination clusters support extended attributes.) If you select one or more of the Preserve options and you are replicating to S3 or ADLS, the values all of these items are saved in metadata files on S3 or ADLS. When you replicate from S3 or ADLS to HDFS, you can select which of these options you want to preserve. **Note:** To preserve permissions to HDFS, you must be running as a superuser on the destination cluster. Use the Run As Username option to set the username.
Delete Policy	Choose the required options to determine whether the files that were deleted on the source should also be deleted from the destination directory. This policy also determines the handling of files in the destination location that are unrelated to the source. Options include: • Keep Deleted Files - Retains the destination files even when they no longer exist at the source. This is the default option. • Delete to Trash - If the HDFS trash is enabled, files are moved to the trash folder. This is not supported when replicating to S3 or ADLS. • Delete Permanently - Uses the least amount of space; use with caution.
Alerts	Choose to generate alerts for various state changes in the replication workflow. You can choose to generate an alert On Failure, On Start, On Success, or On Abort of the replication job. You can configure alerts to be delivered by email or sent as SNMP traps. If alerts are enabled for events, you can search for and view the alerts on the **Events** tab, even if you do not have email notification configured. For example, if you choose Command Result that contains the Failed filter on the Diagnostics Events page, the alerts related to the On Failure alert for all the replication policies for which you have set the alert appear. For more information, see Managing Alerts and Configuring Alert Delivery.

12. Click Create.

What to do next
You can track the replication policy job status on the **Replication Policies** page.

Manage and monitor HDFS replication policies

After you create an HDFS replication policy in Cloudera Replication Manager, you can perform and monitor various tasks related to the replication policy. You can view the job progress and replication logs. You can edit the advanced options to optimize a job run. You can suspend a job and also activate a suspended job. You can edit the replication policy as necessary.

About this task
On the **Replication Policies** page, you can perform the following actions and tasks on a replication policy and its jobs:

Procedure

- When you click Actions for an HDFS replication policy, the following actions appear:

Action	Description
Edit*	Change the replication policy options as required for non-expired policies that are in active or suspended state. Based on the schedule you choose, the replication policy replicates data. You can edit the replication policies to better align with changing requirements. For example, you might want to change the frequency of a policy depending on the data size and importance of the data being replicated. **Note:** A replication policy is associated with a cluster or a cluster pair, therefore you cannot change the clusters in the policy. Optionally, expand a replication policy on the **Replication Policies** page to edit the replication policy options which include frequency (start time cannot be modified if the policy has already started), queue name, maximum bandwidth, and maximum map slots. **Tip:** To optimize the replication policy performance, you can configure the queue name, maximum bandwidth, and maximum map slots as necessary.
Delete	Deletes the replication policy permanently.
Suspend	Suspends a running replication policy. Activate the replication policy, if required.
View Log	Download, copy, or open the log. The log shows a brief output of the stdout and stderr logs of a single step of the latest replication policy job run. You can also view the current job status in the Replication Manager Overview Issues & Updates Job Status column. If the job failed, click **Failed** to view the log details about the job.
Collect diagnostic bundle	Generates a diagnostic bundle for the replication policy. You can download the bundle as a ZIP file to your machine. Ensure that you are logged into the Cloudera Manager instances for both the source and target clusters before you download the bundle in Replication Manager.

*

To view and use the replication policies with an empty name in Replication Manager, you must understand the following implementation:

- If the API version is lower than 51, an existing replication policy with an empty name can be used and updated. However, if you edit the replication policy and provide a name for the replication policy in versions higher or equal to 51, you must ensure that the name conforms to the validation rules.
- If the API version is higher or equal to 51, it is mandatory that you provide a unique name to the replication policy to continue using it. This is because API version 51 and higher enforces the validation rules on all the replication policies.

To pass the replication policy name validation, you must ensure that the replication policy name is unique. The name can contain letters, numbers, and the _ / - characters. You must also ensure that it does not contain the characters % . ; \ nor any character that is not ASCII printable, which includes the ASCII characters less than 32 and the ASCII characters that are greater than or equal to 127.

- When you expand the policy details, the **Job History** panel appears.

  You can view the following details on the panel:

  a) Previous jobs, current job, and one future scheduled job if any.

  b) Job details which include:

Job details	Description
Started	Timestamp when the job started.
Ended	Timestamp when the job ended.
Duration	Time taken to complete the job.

Job details	Description
Progress	Current status of a running job.
Expected	Remaining number of files and bytes expected to be copied for a running job.
Copied	Number of files and bytes copied for a running job and completed job.
Failed	Number of files and bytes that failed to be copied for a completed job.
Deleted	Number of files deleted for a completed job.
Skipped	Remaining number of files and bytes skipped from copying for a running job and complete job.

c) Click Actions to:

- Abort the job.
- Re-run an aborted or failed job.
- View Log for the job. You can download, copy, or open it to track the job and to troubleshoot any issues for the job.

• When you click a job on the **Job History** panel, the following tabs appear:

Tab	Description
General	Shows the following job details: • **Started at** timestamp • **Duration** to complete the job • **HDFS Replication Report** to download the job statistics in CSV format • Job status **Message**
Command Details	Shows the steps that Replication Manager ran for the job along with the timestamp.

• You can download the following CSV reports from the General HDFS Replication Report field to track the replication jobs and to troubleshoot issues:

Report	Description
Listing	Lists all the files and directories copied during the replication job.
Status	Shows the complete status report of each file as: • an **Error** occurred and the file was not copied. • a **Deleted** file. • an up-to-date file for which the replication was **Skipped**.
Error Status	Status report of all the copied files with errors. Each file shows the status, path, and message for the copied files with errors.
Skipped Status	Status report of all skipped files. Each file lists the status, path, and message for the databases and tables that were skipped.
Deleted Status	Status report of all deleted files. Each file lists the status, path, and message for the databases and tables that were deleted.
Performance	Summary report about the performance of the running replication job which includes the last performance sample for each mapper that is working on the replication job.
Full Performance	Performance report of the job which includes the samples taken for all mappers during the replication job.

Note: The reports are generated based on the source Cloudera Manager response. If the Cloudera Manager response is interrupted or is not handled as expected, corresponding error messages appear in HTML format in the reports.

Using Hive replication policies

To create a Hive replication policy in Cloudera Replication Manager, you must configure the required Ranger policy in Ranger, register the on-premises cluster (CDH or Cloudera Base on premises) as a classic cluster in Cloudera Management Console, register cloud account credentials in the Replication Manager service, verify cluster access, and configure minimum ports for replication. The replication load happens on the source on-premises cluster. You can replicate data on-premises to the cloud with a single cluster if the Metastore is running on the cloud.

These policies support table-level replication and can replicate Hive external tables from on-premises clusters (CDH and Cloudera Base on premises) to cloud storage such as S3 and ABFS and to Cloudera Data Hub clusters. They also can:

- replicate data stored in Hive tables, Hive metadata, data in Hive metastore, and Impala metadata (catalog server metadata) associated with Impala tables registered in the Hive metastore, and

  Note: Hive2 managed tables are converted to external tables after replication.

- migrate Sentry permissions to Ranger.

  Note: To perform the Sentry policy replication, you must be running the Sentry service on CDH 5.12 or higher, or any CDH 6.x version.

Hive metadata replication involves multiple entities. Replication Manager supports replication of external tables in Hive. Hive supports replication of external tables to the target cluster and it retains all the properties of external tables. The data files permission and ownership are preserved so that the relevant external processes can continue to write in it even after failover.

You can also use Cloudera CLI commands to create Hive replication policies. The Cloudera CLI commands for Replication Manager are under the replicationmanager Cloudera CLI option. For more information, see *Cloudera CLI for Replication Manager*.

> **Important:** Before you create Hive replication policies, you must ensure that the required Ranger policy is set in Ranger and see Support matrix for Replication Manager on Cloudera on cloud to verify whether your clusters are supported by Replication Manager.

The Apache Ranger access policy model consists of the following components:

- Specification of the resources that you can apply to a replication policy which includes the HDFS files and directories; Hive databases, tables, and columns; and HBase tables, column-families, and columns.
- Specification of access conditions for specific users and groups.

You must set the Ranger policy for the hdfs user on the target cluster to perform all operations on all databases and tables. The same user role is used to import Hive Metastore. The hdfs user should have access to all Hive datasets, including all operations. Otherwise, Hive import fails during the replication process.

On the target cluster, the hive user must have Ranger admin privileges. The same hive user performs the metadata import operation.

For more information about Hive replication policies to replicate data from CDH clusters to Cloudera on cloud, see Migrate Hive data from CDH to Cloudera on cloud blog.

Preparing to create a Hive replication policy

Before you create the Hive replication policies in Cloudera Replication Manager, you must prepare the clusters and verify cluster access and cloud credentials.

Procedure

- Do the source cluster and target cluster meet the requirements to create a Hive replication policy?

  The following image shows the support matrix based on the lowest supported End of Support (EoS) versions that you can use for Hive replication policies. Consult the Support matrix for Cloudera Replication Manager for the complete list of supporting clusters and scenarios:

Figure 4: Lowest supported versions for Hive replication policies

- Is the source CDH cluster or source Cloudera Base on premises cluster registered as a classic cluster on the Management Console?

  CDH clusters and Cloudera Base on premises clusters are managed by Cloudera Manager. To enable these on-premises clusters for Replication Manager, you must register them as Classic Clusters on the Management Console. After registration, you can use them for data migration purposes.

  > ⚠️ **Important:** When you register a Cloudera Base on premises cluster as a classic cluster, ensure that you use only the Cloudera Manager IP Address and Cloudera Manager Port options and not the Register KNOX endpoint (Optional) option.

  For information about registering an on-premises cluster as a classic cluster, see Adding a CDH cluster and Adding a Cloudera Base on premises cluster.

- Does the target Data Hub use Cloudera Manager 7.9.0 or higher? If not, upgrade Cloudera Manager to version 7.9.0 or higher.

- Have you configured the **all-database, table, column** Ranger policy for the hdfs user on the source and target cluster to perform all the operations on all databases and tables?

  The hdfs user role is used to import Hive Metastore and must have access to all Hive datasets, including all operations. Otherwise, Hive import fails during the replication process.

  To provide access, go to the Ranger Admin Web UI Service Manager Hadoop_SQL Policies Access section, and provide hdfs user permission to the **all-database, table, column** policy name.

  Before you select the Additional Settings Include Sentry Permissions with Metadata option during the Hive replication policy creation process, ensure that the *hive* user has the Admin privileges in the Ranger service on the target cluster. This is because the import operation of the Ranger policies is performed by the *hive* user.

  To provide the Admin privileges to the *hive* user, perform the following steps:

  1. Go to the Ranger Admin Web UI.
  2. Click the Settings Users/Groups/Roles option.
  3. Click hive.
  4. Select Admin in the **Select Role** field on the **User Detail** page.
  5. Click Save.

  The **User** list shows *hive* with the Admin role.

  For more information, see Ranger documentation about user editing.

- Is an external account configured on the source CDH cluster's Cloudera Manager which allows the CDH cluster to access Cloudera cloud storage?

  > **Tip:** The external account can be configured on the Cloudera Manager Administration page and the account has the access key / secret key pair that you can use to access Cloudera cloud storage.

- Do you have the required cluster access to create replication policies?

  Power users, the user who onboarded the source and target clusters, and users with ClassicClusterAdmin or ClassicClusterUser resource roles can create replication policies on clusters for which they have access. For more information, see Understanding account roles and resource roles.

  > **Tip:** Ensure that you have *Replication Administrator* or Full Administrator role on the source on-premises cluster.

- Do you have the required cluster access to view the replication policies?

  Existing Hive replication policies are visible to users who have access to the source cluster in the replication policy. A warning appears if you do not have access to the source cluster.

  If you can view the policies, you can perform other actions on the policy including policy update and policy delete operations.

  > **Note:** A Replication Manager user can browse, within Replication Manager, the folder structure of any cluster enabled for Replication Manager. Therefore, permitted users can view folders, files, and databases in Replication Manager that they might not have access to in HDFS. The users cannot view the content of files on the source or destination clusters, nor can they modify or delete folders or files that are viewable from Replication Manager.

- Is the required cloud credential that you want to use in the replication policy registered with the Replication Manager service?

  For more information, see Working with cloud credentials.

- Are the following ports open and available for Replication Manager?

Table 3: Minimum ports required for Hive replication policies

Connectivity required for	Default Port	Type	Description
Data transfer from classic cluster hosts to cloud storage	80 or 443 (TLS)	Outbound	Outgoing port. All classic cluster nodes must be able to access S3/ADLS Gen2 endpoint.
Cloudera Manager Admin Console HTTP	7180 or 7183 (when TLS enabled)	Inbound	Incoming port. Open on the source cluster to enable the target Cloudera Manager in cloud to communicate to the on-premises Cloudera Manager.
Classic cluster	6000-6049 for CCMv1 443 for CCMv2	Outbound	Connecting the source classic cluster to the Cloudera Management Console through Cluster Connectivity Manager (CCM) For more information, see Outbound network access for CCM, and CCM overview.

Consider the following best practices while using on Microsoft Azure ADLS Gen2 (ABFS):

- Ensure that the on-premises cluster (port 443) can access the *https://login.microsoftonline.com* endpoint. This is because the Hadoop client in the on-premises cluster (CDH/) connects to the endpoint to acquire the access tokens before it connects to Azure ADLS storage. For more information, see the *General Azure guidelines* row in the Azure-specific endpoints table.

- Ensure that the steps mentioned in the *General Azure guidelines* and *Azure Data Lake Storage Gen 2* rows in the Azure-specific endpoints table are complete so that the endpoint connects to the target path successfully.

The following system architecture diagram shows the interaction between components during Hive replication using Hive replication policies:

Figure 5: System architecture diagram for Hive replication in Cloudera Replication Manager
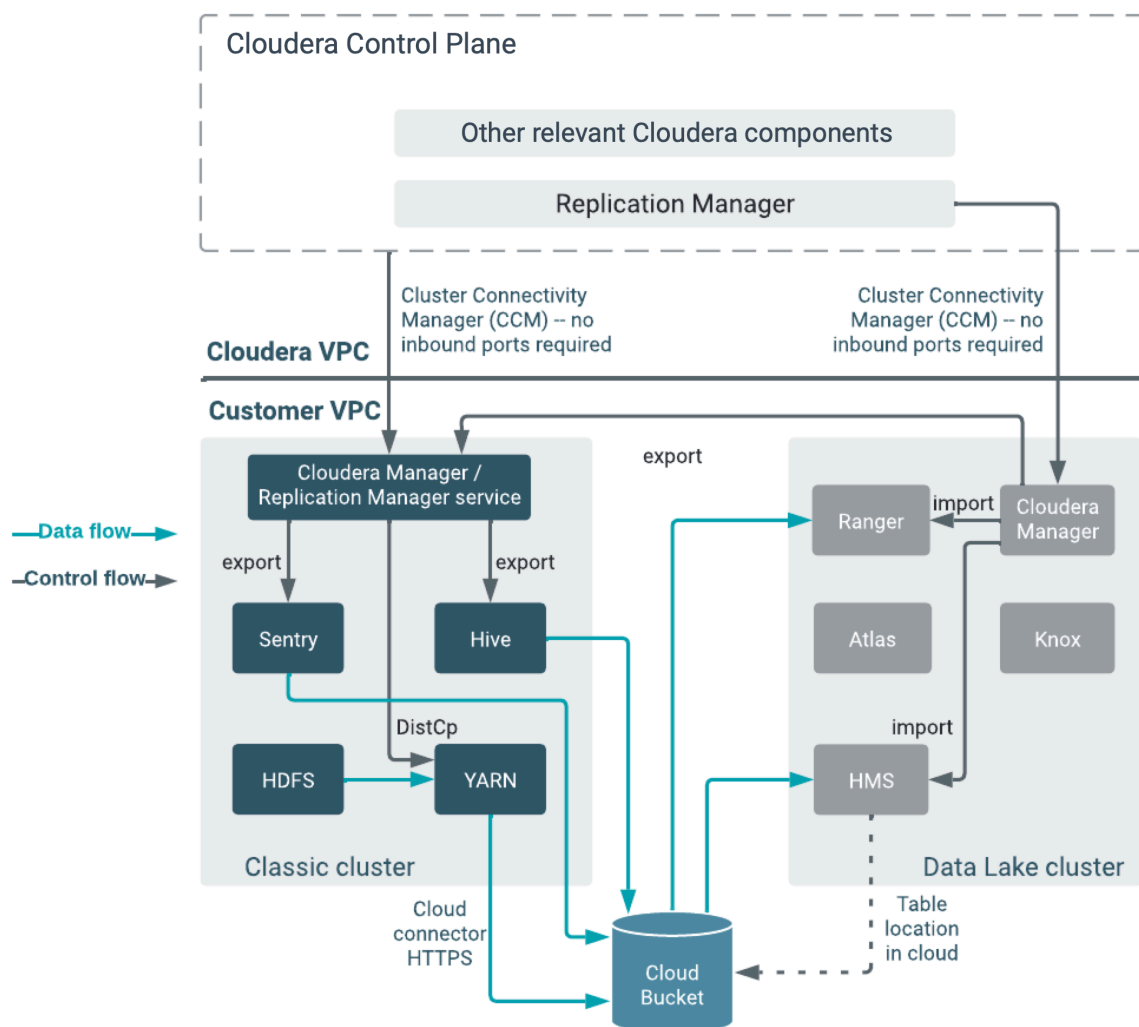
What to do next

After the clusters and cloud storage requirements are met, you can create a Hive replication policy.

Creating Hive replication policy

Before you create a Hive replication policy in Cloudera Replication Manager, you must ensure that the prerequisites are complete and the required Ranger policy is set in Ranger.

About this task

Alternatively, you can also use CDP CLI commands to create Hive replication policies. The CDP CLI commands for Replication Manager are under the replicationmanager CDP CLI option. For more information, see CDP CLI for Replication Manager.

Procedure

1. On the Cloudera Management Console Replication Manager Replication Policies page, click Add Policy.
   The **Create Replication Policy** wizard appears.
2. On the General page, choose or enter the following information:

Option	Description
Hive	Creates a Hive replication policy.
Policy Name	Enter a unique name for the replication policy.
Description	Optional. Enter a brief description about the replication policy.

The following image shows a sample General page in the **Create Replication Policy** wizard:

3. Click Next.
4. On the Select Source page, enter or choose the options as required:

Option	Description
Source cluster	Choose a classic cluster as a source cluster.
Source Databases and Tables	Enter the database name and table name. Click Add to enter more databases and tables as necessary. **Note:** Ensure that you do not add the sys and information_schema databases for replication if you want to choose the Additional Settings Invalidate Impala Metadata on Destination option for the Hive replication policy.
Run As Username (on source)	Optional. The replication policy uses the Default username to replicate data. Enter another username if you want the replication policy to use it to replicate data. Ensure that the user has the necessary permissions to replicate data. **Note:** You must provide a username for Kerberized clusters.

5. Click Next.

---

6. On the Select Destination page, enter or choose the options as required:

Option	Description
Destination Data Lake or Cloudera Data Hub	Choose a Data Lake or Cloudera Data Hub as the destination cluster. The Managed Warehouse Path and the Hive External Table Base Directory Path for the Data Lake appears. **Note:** Administrators can edit the Hive External Table Base Directory field to add another path to override the default storage location for replicated Hive external tables. Before you add another path to override the default storage location, ensure that the following steps are complete in the Ranger UI: - Alter the ranger policy Default: Hive warehouse locations in cm_s3 service to allow the Hive service to access the updated locations of S3 bucket path. - Manually update the Ranger and Sentry permissions. **Note:** The Data Hubs, using Cloudera Manager version 7.9.0 and higher, for which you have access permissions appear automatically. If you want to use a Cloudera Data Hub that appears disabled, you must upgrade the Cloudera Manager version to 7.9.0 or higher for the Data Hub.
Cloud Credential on Source	Choose the required cloud credential. The cloud credentials that you register for the selected source cluster on the Cloud Credentials page appear in this field. You can also add cloud credentials using the Add Cloud Credential link.
Run as Username	Optional. The replication policy uses the Default username to replicate data. Enter another username if you want the replication policy to use it to replicate data.

7. Click Validate Policy.

Replication Manager verifies whether the details provided are correct.

8. Click Next.

9. On the **Schedule** page, choose or enter the following information:

Option	Description
Run Now	Starts to replicate the existing HDFS data after the replication policy creation is complete. Choose the frequency to replicate data periodically.
Schedule Run	Runs the replication policy to replicate data at a later time. Choose the date and time for the first run, and then choose the frequency to replicate data periodically. **Tip:** On the **Replication Policies** page, click to change the timezone.

Option	Description
Frequency	Choose one of the following options: • Does Not Repeat • Custom - In the Custom Recurrence dialog box, choose the time, date, and the frequency to run the policy. Replication Manager ensures that the same number of seconds elapse between the runs. For example, if you choose the Start Time as January 19, 2022 11.06 AM and Interval as 1 day, Replication Manager runs the replication policy for the first time at the specified time in the timezone the replication policy was created in, and then runs it exactly after 1 day that is, after 24 hours or 86400 seconds. **Note:** Ensure that the frequency in a schedule enables a job to finish before the next job starts. Also, ensure that the jobs based on the same policy do not overlap. If a job is not completed before another job starts, the second job does not run and the job status appears as Skipped. If a job is consistently skipped, you might need to modify the frequency of the job.

10. Click Next.

11. On the Additional Settings page, enter or choose the values as necessary:

Option	Description
YARN Queue Name	Enter the name of the YARN queue for the cluster to which the replication job is submitted if you are using Capacity Scheduler queues to limit resource consumption. The default value for this field is default.
Maximum Maps Slots	Set the maximum number of map tasks (simultaneous copies) per replication job. The default value is 20.
Maximum Bandwidth	Adjust this setting so that each map task is throttled to consume only the specified bandwidth. Each map task ((simultaneous copy) is restricted to consume only the specified bandwidth. This is not always exact. The map throttles back its bandwidth consumption during a copy in such a way that the net bandwidth used tends towards the specified value. You can adjust this setting so that each map task is throttled to consume only the specified bandwidth so that the net bandwidth used tends towards the specified value. The default value for the bandwidth is 100MB per second for each mapper.
Number of concurrent HMS connections	Enter the number of concurrent Hive Metastore connections. The connections are used to concurrently import and export metadata from Hive. Increase the number of threads to improve Replication Manager performance. By default, a new replication policy uses 4 connections. • If you set the value to 1 or more, Replication Manager uses multi-threading with the number of connections specified. • If you set the value to 0 or fewer, Replication Manager uses single threading and a single connection.
Sentry Permissions	Choose Include Sentry Permissions with Metadata to migrate Sentry permissions during the replication job. Otherwise, choose Exclude Sentry Permissions from Metadata to not migrate Sentry permissions during the replication job.
Skip URI Privileges	Select this option if you do not want to include URI privileges when you migrate Sentry permissions. During migration, the URI privileges are translated to point to an equivalent location in S3. If the resources have a different location in Amazon S3, do not migrate the URI privileges because the URI privileges might not be valid.

Option	Description
Replication Option	Choose Metadata and Data to replicate metadata and data in files and directories. Otherwise, choose Metadata only to replicate the metadata of files and directories
Directory for metadata file	Enter / or a valid folder path in the target cluster to save the metadata file. If the field is empty or if the specified folder does not exist, Replication Manager creates a new folder.
Force Overwrite	Select to overwrite data in the destination metastore if incompatible changes are detected.
	For example, if the destination metastore was modified, and a new partition was added to a table, this option forces deletion of that partition, overwriting the table with the version found on the source. If you do not choose the option and the Hive replication process detects incompatible changes on the source cluster, Hive replication fails. This sometimes occurs with recurring replications, where the metadata associated with an existing database or table on the source cluster changes over time.
Invalidate Impala Metadata on Destination	Choose the option to run the Impala INVALIDATE METADATA statement per table on the destination cluster after completing the replication. The statement purges the metadata of the replicated tables and views within the destination cluster's Impala upon completion of replication, allowing other Impala clients at the destination to query these tables successfully with accurate results.
	You must run the INVALIDATE METADATA statement manually for the following scenarios:
	• If the destination Cloudera Manager does not have an Impala service. • If the source contains User Defined Functions (UDF).
	Warning: However, this operation is potentially unsafe if DDL operations are being performed on any of the replicated tables or views while the replication is running. In general, directly modifying replicated data/metadata on the destination is not recommended. Ignoring this can lead to unexpected or incorrect behavior of applications and queries using these tables or views.
Replication Strategy	Choose one of the following replication strategies to determine whether the file replication tasks should be distributed among the mappers statically or dynamically.
	• Static distributes file replication tasks among the mappers up front to achieve an uniform distribution based on the file sizes. • Dynamic distributes the file replication tasks in small sets to the mappers, and as each mapper completes its tasks, it dynamically acquires and processes the next unallocated set of tasks.
	The default replication strategy is Dynamic.
MapReduce Service	Choose the MapReduce or YARN service to use.
	Note: When the target cluster of the replication policy is part of a multi-cluster Cloudera Manager, you must ensure that you select the configured YARN service from the same cluster where the replication's target service is deployed.
Log Path	Enter an alternate path for the logs, if required.

Option	Description
Error Handling	Select the following options as necessary: • Skip Checksum Checks - Determines whether to skip checksum checks on the copied files. If selected, checksums are not validated. Checksums are checked by default. **Note:** You must skip checksum checks to prevent replication failure due to non-matching checksums in the following cases: • Replications from an encrypted zone on the source cluster to an encrypted zone on a destination cluster. • Replications from an encryption zone on the source cluster to an unencrypted zone on the destination cluster. • Replications from an unencrypted zone on the source cluster to an encrypted zone on the destination cluster. Checksums are used for two purposes: • To skip replication of files that have already been copied. If Skip Checksum Checks is selected, the replication job skips copying a file if the file lengths and modification times are identical between the source and destination clusters. Otherwise, the job copies the file from the source to the destination. • To redundantly verify the integrity of data. However, checksums are not required to guarantee accurate transfers between clusters. HDFS data transfers are protected by checksums during transfer and storage hardware also uses checksums to ensure that data is accurately stored. These two mechanisms work together to validate the integrity of the copied data. • Skip Listing Checksum Checks - Whether to skip checksum check when comparing two files to determine whether they are same or not. If skipped, the file size and last modified time are used to determine if files are the same or not. Skipping the check improves performance during the mapper phase. Note that if you select the Skip Checksum Checks option, this check is also skipped. • Abort on Error - Whether to abort the job on an error. If selected, files copied up to that point remain on the destination, but no additional files are copied. Abort on Error is not selected by default. • Abort on Snapshot Diff Failures - If a snapshot diff fails during replication, the replication policy uses a complete copy to replicate data. If you select this option, the policy aborts the replication when it encounters an error instead.

Option	Description
Preserve	Choose the required options to preserve the block size, replication count, permissions (including ACLs), and extended attributes (XAttrs) as they exist on the source file system, or to use the settings as configured on the destination file system. By default source system settings are preserved.
	When Permission is selected, and both the source and destination clusters support ACLs, replication preserves ACLs. Otherwise, ACLs are not replicated. When Extended attributes is selected, and both the source and destination clusters support extended attributes, replication preserves them. (This option only displays when both source and destination clusters support extended attributes.)
	If you select one or more of the Preserve options and you are replicating to S3 or ADLS, the values all of these items are saved in metadata files on S3 or ADLS. When you replicate from S3 or ADLS to HDFS, you can select which of these options you want to preserve.
	Note: To preserve permissions to HDFS, you must be running as a superuser on the destination cluster. Use the Run As Username option to set the username.
Delete Policy	Choose the required options to determine whether the files that were deleted on the source should also be deleted from the destination directory. This policy also determines the handling of files in the destination location that are unrelated to the source. Options include:
	• Keep Deleted Files - Retains the destination files even when they no longer exist at the source. This is the default option.
	• Delete to Trash - If the HDFS trash is enabled, files are moved to the trash folder. This is not supported when replicating to S3 or ADLS.
	• Delete Permanently - Uses the least amount of space; use with caution.
	Note: If the source path is globbed, the replication policy ignores the Delete to Trash and Delete Permanently options and performs the Keep Deleted Files operation on the target files. The target files are *not* moved to trash or deleted regardless of the option you choose for the replication policy.
Alerts	Choose to generate alerts for various state changes in the replication workflow. You can choose to generate an alert On Failure, On Start, On Success, or On Abort of the replication job.
	You can configure alerts to be delivered by email or sent as SNMP traps. If alerts are enabled for events, you can search for and view the alerts on the **Events** tab, even if you do not have email notification configured. For example, if you choose Command Result that contains the Failed filter on the Diagnostics Events page, the alerts related to the On Failure alert for all the replication policies for which you have set the alert appear. For more information, see Managing Alerts and Configuring Alert Delivery.

12. Click Create.

What to do next

After the replication policy runs successfully, you can view the replication job status on the **Replication Policies** page. Verify whether the job starts and runs as expected.

The administrator can verify whether the job starts and runs as expected using the following steps:

- For HDFS data replication, check the cloud storage path (for example, S3 bucket path) to verify whether the data was successfully copied in the specified bucket.
- For Hive metadata replication, verify whether the specified source database, along with tables, partitions, UDFs and column stats are available in the Data Lake HMS instance. For this, the administrator can use a Cloudera Data Hub cluster and run the corresponding queries using Hue or beeline.

- For Ranger policies, query the Ranger policies to ensure that the Sentry policies are properly mapped as Ranger policies for the right users and groups.

  Note: If the CDH source database contains functions, you must explicitly run the `reload function` command to view the migrated replication functions in the target location.

Manage and monitor Hive replication policies

After you create a Hive replication policy in Cloudera Replication Manager, you can perform and monitor various tasks related to the replication policy. You can view the job progress and replication logs. You can edit the advanced options to optimize a job run. You can suspend a job and also activate a suspended job. You can edit the replication policy as necessary.

About this task

On the **Replication Policies** page, you can perform the following actions and tasks on a replication policy and its jobs.

Procedure

- When you click Actions for a Hive replication policy, the following actions appear:

Action	Description
Edit*	Change the replication policy options as required for non-expired policies that are in active or suspended state. Based on the schedule you choose, the replication policy replicates data.
	You can edit the replication policies to better align with changing requirements. For example, you might want to change the frequency of a policy depending on the data size and importance of the data being replicated.
	Note: A replication policy is associated with a cluster or a cluster pair, therefore you cannot change the clusters in the policy.
	Optionally, expand a replication policy on the **Replication Policies** page to edit the replication policy options which include frequency (start time cannot be modified if the policy has already started), queue name, maximum bandwidth, and maximum map slots.
	Tip: To optimize the replication policy performance, you can configure the queue name, maximum bandwidth, and maximum map slots as necessary.
Delete	Deletes the replication policy permanently.
Suspend	Suspends a running replication policy. Activate the replication policy, if required.
View Log	Download, copy, or open the log. The log shows a brief output of the stdout and stderr logs of a single step of the latest replication policy job run.
	You can also view the current job status in the Replication Manager Overview Issues & Updates Job Status column. If the job failed, click **Failed** to view the log details about the job.
View Command Details	Opens the latest Hive replication policy job page. The steps and substeps appear in a tree view. The failed steps are expanded by default, showing the last 15 lines of the log.
	You can also view the command details for a Hive replication policy on the Overview Issues & Updates panel.
	Tip: To view the complete log for all the jobs, go to the target cluster Cloudera Manager Running Commands page.

Action	Description
Collect diagnostic bundle	Generates a diagnostic bundle for the replication policy. You can download the bundle as a ZIP file to your machine. Ensure that you are logged into the Cloudera Manager instances for both the source and target clusters before you download the bundle in Replication Manager.

*

To view and use the replication policies with an empty name in Replication Manager, you must understand the following implementation:

- If the API version is lower than 51, an existing replication policy with an empty name can be used and updated. However, if you edit the replication policy and provide a name for the replication policy in versions higher or equal to 51, you must ensure that the name conforms to the validation rules.
- If the API version is higher or equal to 51, it is mandatory that you provide a unique name to the replication policy to continue using it. This is because API version 51 and higher enforces the validation rules on all the replication policies.

To pass the replication policy name validation, you must ensure that the replication policy name is unique. The name can contain letters, numbers, and the _ / - characters. You must also ensure that it does not contain the characters % . ; \ nor any character that is not ASCII printable, which includes the ASCII characters less than 32 and the ASCII characters that are greater than or equal to 127.

- When you expand the policy details, the **Job History** panel appears.

  You can view the following details on the panel:

  a) Previous jobs, current job, and one future scheduled job if any.
  b) Job details which include:

Job details	Description
Started	Timestamp when the job started.
Ended	Timestamp when the job ended.
Duration	Time taken to complete the job.
Tables	Number of imported or exported tables.
Progress	Current status of a running job.
Expected	Remaining number of files and bytes expected to be copied for a running job.
Copied	Number of files and bytes copied for a running job and completed job.
Failed	Number of files and bytes that failed to be copied for a completed job.
Deleted	Number of files deleted for a completed job.
Skipped	Remaining number of files and bytes skipped from copying for a running job and complete job.

  c) Click Actions to:

  - Abort the job.
  - Re-run an aborted or failed job.
  - View Log for the job. You can download, copy, or open it to track the job and to troubleshoot any issues for the job.

• When you click a job on the **Job History** panel, the following tabs appear:

Tab	Description
General	Shows the following job details: • **Started at** timestamp • **Duration** taken to complete the job • **HDFS Replication Report** to download the job statistics in CSV format • **Hive Replication Report** to download the job statistics in CSV format • **Hive Export/Import** is the number of external tables exported or imported using Hive replication. • Number of **Errors** encountered during the replication job. • **Impala UDFs** is the number of tables exported or imported using Impala. • Job status **Message**.
Command Details	Shows the details about the commands that ran on the source Cloudera Manager for the job, along with the timestamp.
Setup Error	Shows the stack trace for the commands that ran on the source Cloudera Manager for the failed job.

• You can download the following CSV reports from the General HDFS Replication Report field to track the replication jobs and to troubleshoot issues:

Report	Description
Listing	Lists all the files and directories copied during the replication job.
Status	Shows the complete status report of each file as: • an **Error** occurred and the file was not copied. • a **Deleted** file. • an up-to-date file for which the replication was **Skipped**.
Error Status	Status report of all the copied files with errors. Each file shows the status, path, and message for the copied files with errors.
Skipped Status	Status report of all skipped files. Each file lists the status, path, and message for the databases and tables that were skipped.
Deleted Status	Status report of all deleted files. Each file lists the status, path, and message for the databases and tables that were deleted.
Performance	Summary report about the performance of the running replication job which includes the last performance sample for each mapper that is working on the replication job.
Full Performance	Performance report of the job which includes the samples taken for all mappers during the replication job.

> **Note:** The reports are generated based on the source Cloudera Manager response. If the Cloudera Manager response is interrupted or is not handled as expected, corresponding error messages appear in HTML format in the reports.

• You can download the following CSV reports from the General Hive Replication Report field to track the replication jobs and to troubleshoot issues:

Report	Description
Hive Result	List of replicated tables.
Hive Performance	Performance report for Hive replication.

Using HBase replication policies

To create an HBase replication policy in Cloudera Replication Manager, you must register the on-premises cluster (CDH or Cloudera Base on premises) as a classic cluster in Cloudera Management Console, register cloud account credentials in the Replication Manager service, verify cluster access, and configure minimum ports for replication.

Preparing to create an HBase replication policy

Before you create HBase replication policies in Cloudera Replication Manager, you must prepare the clusters, register cloud storage in Replication Manager, and verify cluster access.

Procedure

- Do the source cluster and target cluster meet the requirements to create an HBase replication policy?

  The following image shows the support matrix based on the lowest supported End of Support (EoS) versions that you can use for HBase replication policies. Consult the Support matrix for Cloudera Replication Manager for the complete list of supporting clusters and scenarios:

  Figure 6: Lowest supported versions for HBase replication policies

- Is the source CDH cluster or source Cloudera Base on premises cluster registered as a classic cluster on the Management Console?

  CDH clusters and Cloudera Base on premises clusters are managed by Cloudera Manager. To enable these on-premises clusters for Replication Manager, you must register them as Classic Clusters on the Management Console. After registration, you can use them for data migration purposes.

  > ⚠️ **Important:** When you register a Cloudera Base on premises cluster as a classic cluster, ensure that you use only the Cloudera Manager IP Address and Cloudera Manager Port options and not the Register KNOX endpoint (Optional) option.

  For information about registering an on-premises cluster as a classic cluster, see Adding a CDH cluster and Adding a Cloudera Base on premises cluster.

  > 📝 **Note:**
  > - Cloudera Base on premises 7.1.6 and higher clusters must be Kerberos enabled to use them as source Classic Clusters.
  > - You must have *Full Administrator* role on the source on-premises cluster.

- Are the following steps complete on the source Cloudera Base on premises cluster or source CDH cluster (these steps are not required for COD sources)?

  1. Complete Step 1 of Migrating HBase data from CDH/HDP to COD to install the HBase replication plugin parcel in the CDH source clusters.

     This step is applicable for CDH versions 7.2.x that are lower than 7.2.2, versions 7.1.x that are lower than 7.1.5, and for versions lower than 7.x.

  2. Create the /user/hbase folder for the hbase user in HDFS in the source cluster using the following commands:

     ```
     sudo -u hdfs hdfs dfs -mkdir /user/hbase
     sudo -u hdfs hdfs dfs -chown hbase:hbase /user/hbase
     ```

     These commands allow the HBase replication policy to replicate the existing data in the source cluster.

     This step is applicable for Cloudera Manager versions 7.4.3 or lower; Cloudera Manager 7.4.4 only if the API version is lower than v45. The endpoint http://*[***CM_HOST***]:[***CM_PORT***]*/api/version shows the API version of the Cloudera Manager.

- Is the required target cluster (Data Hub or COD) available and healthy?

  Note: HBase replication policies do not support source or destination Cloudera on cloud clusters created using the Micro Duty template.

- Do you have the necessary permission to run the HBase replication jobs on YARN?

  To verify whether you have the necessary permission, perform the following steps:

  1. Go to the source Cloudera Manager *YARN SERVICE* Configuration tab.
  2. Ensure one of the following conditions is met:

     - The Allowed System Users property must have *hbase*. Otherwise, add hbase to the existing property value.

       This property lists the users permitted (or allowed) to run containers. Note that the users with IDs lower than the Minimum User ID property might be permitted (or allowed) to run containers.
     - The Minimum User ID property is set to a value that is lower than the HBase user's ID.

       To check the HBase user's ID, SSH into a cluster node and run the id hbase command.

       The following sample snippet shows the HBase user's ID when you run the id hbase command:

       ```
       # id hbase
       uid=39993(hbase) gid=39993(hbase) groups=39993(hbase)
       ```

- Is the required cloud credential that you want to use in the replication policy registered with the Replication Manager service?

  For more information, see Working with cloud credentials.

  You can also add the following advanced configuration settings to use Google Cloud, Amazon S3, and ADLS accounts in Replication Manager:

  1. Go to the source Cloudera Manager Clusters *HDFS SERVICE* Configuration tab.
  2. Locate the HDFS Client Advanced Configuration Snippet (Safety Valve) for hdfs-site.xml property.
  3. Add the following key-value pairs to register a Google account to use in Replication Manager:

     - fs.gs.impl=com.google.cloud.hadoop.fs.gcs.GoogleHadoopFileSystem
     - fs.gs.project.id=*[***ENTER THE PROJECT ID***]*
     - fs.gs.system.bucket=*[***ENTER THE BUCKET NAME***]*
     - fs.gs.working.dir=/
     - fs.gs.auth.service.account.enable=true
     - fs.gs.auth.service.account.email=*[***ENTER THE SERVICE PRINCIPAL EMAIL ID***]*
     - fs.gs.auth.service.account.keyfile=/*[***ENTER THE LOCAL PATH OF THE P12 FILE***]*
  4. Add the following key-value pairs to register an S3 account to use in Replication Manager:

     - fs.s3a.access.key=*[***ENTER THE SESSION ACCESS KEY***]*
     - fs.s3a.secret.key=*[***ENTER THE SESSION SECRET KEY***]*
  5. Add the following key-value pairs to register an ADLS account to use in Replication Manager:

     - fs.azure.account.oauth2.client.id=*[***ENTER THE ABFS STORAGE CLIENT ID***]*
     - fs.azure.account.oauth2.client.secret=*[***ENTER THE ABFS STORAGE CLIENT SECRET KEY***]*
  6. Save and restart the HDFS service for the changes to take effect.
- Do you have the required cluster access to create or view replication policies?
- Have you assigned the managed identity of source roles, Storage Blob Data Owner or Storage Blob Data Contributor, to the destination storage data container and vice versa for bidirectional replication when you are using COD on Microsoft Azure?

  The roles allow writing a snapshot in the destination cluster container.
- Does DNS resolution work as expected between the source and destination clusters?

  Tip: If the destination cluster is not reachable from the source RegionServer hosts, add the hostname and IP address of the destination hosts to the /etc/hosts file on the RegionServers of the source cluster.

- Is the outgoing SSH port open on the Cloudera Manager host?
- Are the following ports open and available for Replication Manager?

Table 4: Minimum ports required for HBase replication policies

Ports	Service	Description
2181 and 16020	Destination hosts of the AWS cluster or ADLS cluster (target cluster), and the Cloudera Manager server port on the source cluster	Verify whether the ports 16020 for worker security group and 2181 for worker, master, and leader groups are open for connection from the source cluster to the destination cluster on AWS or Azure. This ensures that the source HBase service can communicate with Zookeeper and HBase services on the destination hosts uninterruptedly. For more information, see Ports for HBase replication.

Ports	Service	Description
16000	HMaster	Open the port on the Master Nodes (HBase Master Node and any back-up HBase Master node). Before you select the Validate Replication option during the first HBase replication policy creation between two specific clusters, you must ensure that the port is open on the target cluster. **Note:** Irrespective of whether this port is open or not on the Master nodes, Replication Manager displays a warning message to inform you that this port should be open on the target cluster (to communicate with the source cluster) when you choose Validate Replication on the **Select Destination** page during the HBase replication policy creation process.
7180 or 7183	Cloudera Manager Admin Console HTTP	Open on the source cluster to enable Data lake Cloudera Manager to communicate to the on-premises Cloudera Manager. Connects to destination SDX Data Lake Cloudera Manager.
9000	Cloudera Manager Agent	Open on the source and target cluster to retrieve diagnostic and log information.
6000-6049	Cluster Connectivity Manager (CCM)	Required for SSL connections to the Control Plane via CCM to communicate with Replication Manager.
80 or 443	Data transfer from secondary node for AWS / ADLS Gen2	Outgoing port. Open on all the HDFS nodes for AWS and ADLS Gen2.
8443	Data Lake cluster	Outgoing port. Configure the port on the Data Lake cluster as the outgoing port for Cloudera Management Console to communicate with Cloudera Manager and Knox.
8032	YARN Resource Manager	Open on the source and target cluster to access the YARN ResourceManager.

Consider the following best practices while using on Microsoft Azure ADLS Gen2 (ABFS):

- Ensure that the on-premises cluster (port 443) can access the *https://login.microsoftonline.com* endpoint. This is because the Hadoop client in the on-premises cluster (CDH/) connects to the endpoint to acquire the access tokens before it connects to Azure ADLS storage. For more information, see the *General Azure guidelines* row in the Azure-specific endpoints table.
- Ensure that the steps mentioned in the *General Azure guidelines* and *Azure Data Lake Storage Gen 2* rows in the Azure-specific endpoints table are complete so that the endpoint connects to the target path successfully.

- Are the following steps complete to use the Export snapshot user field during the HBase replication policy creation process?

  Replication Manager uses the specified username on the source cluster to export the initial snapshots to the target. The *hbase* user impersonates the specified *Export snapshot user* credentials and permissions to submit the HBase YARN application to export the HBase snapshot.

  To ensure the YARN application succeeds, the *hbase* user must be allowed to impersonate the *exportSnapshotUser* in HDFS by configuring the required properties and values in the HDFS core_site_safety_valve of the source cluster.

  For example, if you want to allow the impersonation from any host and if the *exportSnapshotUser* is in the *repl* user group, you can set the following key-value pairs in source Cloudera Manager Clusters *CORE SETTINGS* Configuration Cluster-wide Advanced Configuration Snippet (Safety Valve) for core-site.xml :

  - hadoop.proxyuser.hbase.groups = repl
  - hadoop.proxyuser.hbase.hosts = *

  The Export snapshot user field appears in the **Create HBase replication policy** wizard when the following conditions are true:

  - The target cluster version is 7.2.18.200 and higher.
  - The source cluster version is 7.1.9 SP1 using Cloudera Manager 7.11.3 CHF7 and higher.
  - You choose the Select Source Perform Initial Snapshot option in the **Create HBase replication policy** wizard.

  When you use IDBroker credentials to replicate from Cloudera Base on premises clusters, you must map the Kerberos username to AWS in the *[***USERNAME***]=[***ARN***]* format in the source Cloudera Manager Clusters *KNOX SERVICE* Instances Configuration Knox IDBroker AWS User Mapping property, and allow the *hbase* user to impersonate this user in IDBroker.

- (Optional) Complete the steps mentioned in the Optimize HBase replication policy performance when replicating HBase tables with several TB data FAQ if you choose Perform Initial Snapshot during HBase replication policy creation to replicate HBase tables with several TB data.

What to do next

After the clusters and cloud storage requirements are met, you can create an HBase replication policy.

Creating HBase replication policy

You can replicate HBase data from a source classic cluster (CDH or Cloudera Base on premises cluster), Cloudera Operational Database, or Cloudera Data Hub to a target Cloudera Data Hub or Cloudera Operational Database cluster in Cloudera Replication Manager.

Before you begin

Ensure that the HBase replication policy prerequisites are complete.

Procedure

1. On the Cloudera Management Console Replication Manager Replication Policies page, click Add Policy.

   The **Create Replication Policy** wizard appears.

2. On the General page, choose or enter the following information:

Option	Description
HBase	Creates an HBase replication policy.
Policy Name	Enter a unique name for the replication policy.

Option	Description
Description	Optional. Enter a brief description about the replication policy.

The following image shows a sample General page in the **Create Replication Policy** wizard:
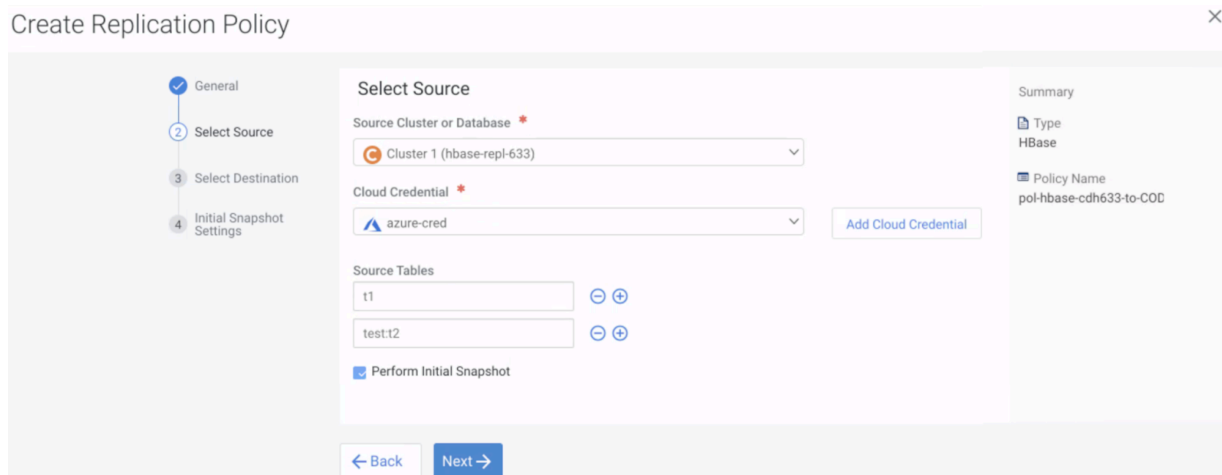
3. Click Next.
4. On the Select Source page, enter or choose the options as required:

Option	Action
Source Cluster or Database	Choose a source cluster. **Note:** HBase replication policies do not support source Cloudera on cloud clusters created using the Micro Duty template.
Source Tables	Enter a table name that you want to replicate. Click the Add icon to add more table names.
Perform Initial Snapshot	Select the option to replicate existing data.
Credentials are available in source cluster HDFS service configuration setting	You can choose this option when you want to use a Google Cloud account. You can use this option for S3 and ADLS accounts as well. Before you use this option, ensure that the advanced configuration settings in Preparing to create an HBase replication policy are configured.
Credentials from External Account	Choose this option for S3 and ADLS storage options. This option appears when you choose a Cloudera Base on premises cluster or CDH cluster as the source cluster and you choose Perform Initial Snapshot. Click Add Cloud Credential. In the Add Cloud Credential dialog box, enter a unique name for the cloud credential. Click Save after you choose one of the following cloud storage types and enter the required options: • S3 - Choose an authentication type, enter an access key and secret key. • ADLS - Enter the client ID, tenant ID, and secret key.

Option	Action
Replicate Database	Replication Manager replicates: • current and future data from the existing tables if you choose the Perform Initial Snapshot option. Otherwise, only the future data is replicated. • data from the future tables that are created after policy creation. To replicate data from the future tables successfully, you must create similar empty tables on the target cluster. You can perform this action when you create or add a table to the database on the source cluster. You can choose the Replicate Database option only if the following conditions are true: • Target Cloudera Manager version is 7.11.0 or higher. • Source cluster version is CDH 6.x or higher. CDH 5.16.2 and higher versions also support the Replicate Database option after you upgrade the source cluster Cloudera Manager. • No existing HBase replication policies exist between the source and target clusters. **Tip:** If you want to replicate the new tables that are created after the replication policy creation is complete, you must configure the replication scope to "1" for those tables on the source cluster. To configure the replication scope for a table on the master cluster, run the alter *[***TABLE NAME***]*, {NAME => *[***COLUMN FAMILY***]*, REPLICATION_SCOPE => 1} command for each column family that must be replicated. REPLICATION_SCOPE is a column-family level attribute, where the value '0' means replication is disabled, and '1' means replication is enabled.
Replicate all user tables	Appears after you choose the Replicate Database option. Choose this option to replicate all the HBase tables in the database. This action sets the replication scope to 1 for all the tables and then replicates the tables to the target cluster.
Replicate only tables where replication is already enabled	Appears after you choose the Replicate Database option. Choose this option to replicate only those tables for which the replication scope is already set to 1. This provides you a choice to replicate only the required tables in a database. This option is supported only if the target cluster Cloudera version is 7.2.17.300 using Cloudera Manager 7.11.0-h3 or higher versions, Cloudera 7.2.16.500 using Cloudera Manager 7.9.0-h7 or higher versions, or Cloudera 7.12.0.0. To enable this feature, contact your Cloudera Account team.

Option	Action
Export snapshot user	Enter the custom username. Replication Manager uses the specified username to export the initial snapshots to the target. The option appears after you choose the Select Source Perform Initial Snapshot option. **Note:** Before you specify the username, complete the prerequisites for this field as shown in HBase replication policy prerequisites. To enable this feature, contact your Cloudera Account team.

The following sample image shows the Select Source page in the Create Replication Policy wizard when you choose a COD as source cluster:

5. Click Next.
6. On the Select Destination page, enter or choose the options as required:

Option	Description
Destination Cloudera Data Hub or Cloudera Operational Database	Choose a Cloudera Data Hub cluster or Cloudera Operational Database. **Note:** HBase replication policies do not support destination Cloudera on cloud clusters created using the Micro Duty template.
Set HBase Replication Machine User	Optional. Choose the option and then enter the username and password. Ensure that you enter the correct password for an existing user because if the password is incorrect, the data is not replicated even though the policy is created successfully. Based on the username and password that you enter, one of the following possible scenarios is implemented by Replication Manager: • If Set HBase Replication Machine User is not selected, an HBase replication machine user is created automatically with an auto-generated username. • If Set HBase Replication Machine User is selected and Create User If Does Not Exist is not selected, ensure that the username you enter exists in the Cloudera User Management System (UMS), otherwise an error message appears. • If Set HBase Replication Machine User is selected and Create User If Does Not Exist is selected and the username does not exist in UMS, the username is created.

Option	Description
Sync Replication User	Optional. Replication Manager validates the existing username with the UMS and synchronizes the new username and password to the destination cluster's environment (and to the source's as well if the source is Cloudera Operational Database). **Note:** Error appears after you click Sync Replication User if you entered a non-existent username and did not choose Create User If Does Not Exist.
Replicate via a Network Load Balancer	Enable if the source on-premises cluster uses a network load balancer (NLB) to communicate with ZooKeeper and RegionServers of the destination Cloudera Manager of Cloudera Operational Database clusters. You can use this option when the Cloudera Operational Database clusters are isolated and the on-premises clusters can only use NLB to communicate with them. The option is disabled by default if the chosen source and target clusters already have HBase replication policies between them. To enable this feature, contact your Cloudera Account team.
Endpoint	Enter the endpoint details in the *[***NLB ADDRESS***]:[***ZOOKEEPER LISTENER PORT (2181)***]:[***TARGET HBASE ZNODE PATH***]* format. For example, my.nlb.us-west-2.amazonaws.com:2181:/hbase.

The following options appear if the source Cloudera Manager is 7.6.0 or higher:

Option	Description
Rolling HBase Service Restart on Source	[Appears if you select Cloudera Operational Database or Cloudera Data Hub as the source cluster] Select to enable automatic rolling restart* of HBase service on the source cluster after the HBase replication policy first-time setup steps are complete. Otherwise, Cloudera Manager performs an automatic full restart* of the service.
Rolling HBase Service Restart on Destination	Select to enable automatic rolling restart* of HBase service on the target cluster as a rolling restart* after the HBase replication policy first-time setup steps are complete. Otherwise, Cloudera Manager performs an automatic full restart of the service.
I want to force the setup of this HBase replication policy	Choose to run the first-time setup configuration between the selected source and destination clusters. This option appears when the selected source or target cluster is part of an existing cluster pair, and one of the following is true about the cluster pair: • No HBase replication policies exist between them. • The other cluster in the pair is currently unreachable. When you select the option, you acknowledge that the existing pairing for the selected source or target cluster will be cleared and the first-time setup will be initiated with the chosen new source or destination cluster.

Option	Description
Validate Replication	Select the option to notify Replication Manager to verify the provided details so that the replication is initiated after the policy creation is complete.
	Note: This option is available for target clusters using Cloudera Manager version 7.6.2 and higher.
	Before you select the Validate Replication option during the first HBase replication policy creation between two specific clusters, you must ensure that the 16000 port is open on the target cluster.
	Note: Irrespective of whether this port is open or not on the Master nodes, Replication Manager displays a warning message to inform you that this port should be open on the target cluster (to communicate with the source cluster) when you choose Validate Replication on the **Select Destination** page during the HBase replication policy creation process.

*During rolling restart, one node is restarted at a time and this continues until all the nodes in the cluster are restarted. This type of restart ensures that there is no disruption of service. During full restart, all the nodes are shut down at once and restarted simultaneously.

The following sample image shows the Select Destination page in the Create Replication Policy wizard:

7. Click Next.
8. On the Initial Snapshot Settings page, configure the following options for the source cluster:

Option	Description
YARN Queue Name	Enter the name of the YARN queue for the cluster to which the replication job is submitted only if you are using Capacity Scheduler queues to limit resource consumption. The default value for this field is default.
Maximum Maps Slots	Configure the maximum number of map tasks (simultaneous copies) per replication job. The default value is 20.

Option	Description
Maximum Bandwidth	Adjust this setting so that each map task is throttled to consume only the specified bandwidth.
	Each map task ((simultaneous copy) is restricted to consume only the specified bandwidth. This is not always exact. The map throttles back its bandwidth consumption during a copy in such a way that the net bandwidth used tends towards the specified value. You can adjust this setting so that each map task is throttled to consume only the specified bandwidth so that the net bandwidth used tends towards the specified value. The default value for the bandwidth is 100MB per second for each mapper.
	You can adjust the setting only if the source and destination Cloudera Manager instances support this option.
	To enable this feature, contact your Cloudera Account team.
Maximum parallel snapshots	Specify the maximum number of tables to process in parallel during the initial snapshot export and import step for the HBase replication policy.
	If you do not enter any value, Replication Manager chooses an appropriate value, depending on the resources in the source and target cluster, to optimize the performance.
	To enable this feature, contact your Cloudera Account team.

The following sample image shows the Initial Snapshot Settings page in the Create Replication Policy wizard:

9. Click Create.

10. Restart the HBase service on the on-premises source cluster when the policy status on the **Replication Policies** page shows Manual restart (src) / restarting (dest) or Manual HBase restart needed on source. After the service restart is complete, the setup continues automatically for the replication policy. You do not need to restart the HBase service if the source is Cloudera Operational Database or Cloudera Data Hub.

> ⚠️ **Important:** If the source cluster Cloudera Manager version is 7.6.0 or lower and you are using an on-premises source cluster, you must perform the following steps to complete the HBase replication policy setup:
>
> **a.** Restart the HBase service on the on-premises source cluster when the policy status on the **Replication Policies** page shows **Waiting for 'Continue Setup' action call**.
> **b.** Click Continue setup for the policy on the **Replication Policies** page after the service restart is complete. This action informs Replication Manager to continue the replication policy setup.

The following image shows the Continue setup option for the HBase replication policy on the Replication Policies page:
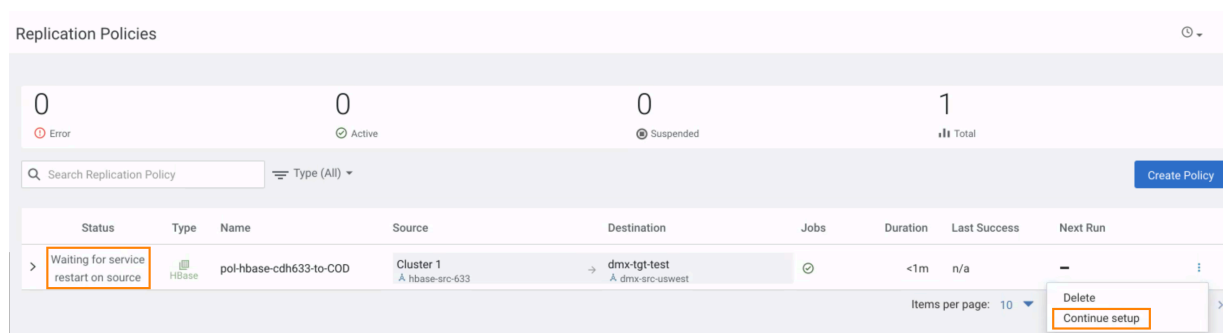
Results

After you create the first replication policy between a source cluster and target cluster (policy that is in setup/service restart state), Replication Manager creates and runs two schedules or jobs. The first schedule shows the service configuration and service restart progress, and the second schedule shows the policy creation progress. Subsequent replication policies between the same source cluster and target cluster creates only one job. Replication Manager restarts the HBase services on both the clusters if they are Cloudera Operational Database clusters.

> 📝 **Note:** The first HBase replication policy between a source cluster and target cluster takes more time to complete because of the first-time setup configuration process that runs in the background. You can continue creating more HBase replication policies while the setup process of the in-progress policies is running in the background. The subsequent replication policies for the same source and target cluster do not run the first-time setup configuration.

What to do next

To verify whether a replication policy is running, you can either click the replication policy on the Replication Policies page, or click Running Commands in Cloudera Manager.

The following sample image shows the Job History page of a HBase replication policy between Cloudera Operational Database clusters:

Manage and monitor HBase replication policies

After you create an HBase replication policy in Cloudera Replication Manager, you can perform and monitor various tasks related to the replication policy. You can also view the job progress and replication logs.

About this task

You can perform the following actions and tasks on the replication policy and its jobs on the **Replication Policies** page:

Procedure

- You can edit the policy name and description; view the source cluster name and tables selected in it to be replicated; and view the destination Data Hub or COD and the tables that have been replicated. You cannot remove the tables if you choose Select Source Replicate Database option during the HBase replication policy creation process.
- On the **Job History** tab, you can view the previous replication jobs, current running job, and one future scheduled job for the replication policy. You can perform **Actions** on the HBase replication policy.
- On the **Charts** tab, you can view the HBase RegionServer metrics for a specific replication peer.

Monitor HBase replication policy job details

When you click a replication policy on the "Replication Policies" page, the "Job History" tab appears. On this tab, you can view the previous replication jobs, current running job, and one future scheduled job for the replication policy.

Replication policy job details

The following columns appear for each replication policy:

- **Type** of job. Click the job to view the completed and in-progress steps that Replication Manager takes to complete the job. Here, you can verify whether the first-time setup configuration between the source and destination clusters was successful.

  > **Tip:** You can also view the progress and results of the first-time setup on the Cloudera Manager Running Commands page for the source and destination clusters.

- Timestamp when the job **Started**.
- Timestamp when the job **Ended**.
- **Duration** taken to complete the job.
- **Progress** of the job run.

Actions menu

On the **Actions** menu, you can perform the following tasks on the HBase replication policy:

Option	Description
Edit*	Enables you to edit the replication policy name and description, or delete one or more tables for the HBase replication policy. You can also edit **Source** to delete the tables in the HBase replication policy.
Delete	Removes the HBase replication policy permanently. You can Force Delete an HBase replication policy after Replication Manager fails to delete the replication policy.
Suspend	Pauses an active HBase replication policy. This option appears for target clusters using Cloudera Manager 7.9.0 (Cloudera Manager API** version 51) and higher. ⚠️ **Warning:** When you suspend an HBase replication policy, Replication Manager suspends all the HBase replication policies between the same source and destination cluster. You can suspend only the specified replication policy for target clusters using Cloudera Manager with API version lower than 45. However, the Suspend option is not available for clusters using Cloudera Manager with API versions 45 through 50.
Activate	Resumes a suspended HBase replication policy. ⚠️ **Warning:** When you activate a suspended HBase replication policy, Replication Manager activates all the suspended HBase replication policies between the same source and destination cluster. You can activate only the specified suspended policy for target clusters using Cloudera Manager with API** version lower than 51.
View command details	Opens the latest HBase replication policy job page. The steps and substeps appear in a tree view. The failed steps are expanded by default, showing the last 15 lines of the log. You can also view the command details for a Hive replication policy on the Overview Issues & Updates panel. To view the complete log for all the jobs, go to the target cluster Cloudera Manager Running Commands page.

Option	Description
Retry First Time Setup	Runs the first-time setup configuration between the source and destination clusters if the first-time setup has failed. This option is available only if the first-time setup configuration fails.
	After you click Retry Create, you can choose:
	• to set an Set HBase Replication Machine User and Sync Replication User to run the first-time setup.
	• to opt for Rolling HBase Service Restart on Source and Rolling HBase Service Restart On Destination after the first-time setup is complete.
	For more information about these options, see *Creating HBase replication policy*.
Retry Failed Snapshots	Reruns the failed initial snapshots (and only the failed ones) in the replication policy if the Replication Manager failed to replicate the existing data of some tables.
	This option appears if you selected Select Source Perform Initial Snapshot during policy creation and Replication Manager failed to replicate the existing data. In this scenario, the policy status shows **Active** with a *Snapshot Failure* warning message on the **Replication Policies** page for the HBase replication policy.
Retry Failed Create	Retries the HBase replication policy creation process. This option appears for failed HBase replication policies that show the **Create failed** status.
Collect diagnostic bundle	Generates a diagnostic bundle for the replication policy. You can download the bundle as a ZIP file to your machine.
	Ensure that you are logged into the Cloudera Manager instances for both the source and target clusters before you download the bundle in Replication Manager.

*

To view and use the replication policies with an empty name in Replication Manager, you must understand the following implementation:

• If the API version is lower than 51, an existing replication policy with an empty name can be used and updated. However, if you edit the replication policy and provide a name for the replication policy in versions higher or equal to 51, you must ensure that the name conforms to the validation rules.
• If the API version is higher or equal to 51, it is mandatory that you provide a unique name to the replication policy to continue using it. This is because API version 51 and higher enforces the validation rules on all the replication policies.

To pass the replication policy name validation, you must ensure that the replication policy name is unique. The name can contain letters, numbers, and the _ / - characters. You must also ensure that it does not contain the characters % . ; \ nor any character that is not ASCII printable, which includes the ASCII characters less than 32 and the ASCII characters that are greater than or equal to 127.

**

The endpoint http://*[***CM_HOST***]*:*[***CM_PORT***]*/api/version shows the API version of the .

Creating triggers and monitoring replication-related metrics in Cloudera Manager

After you create an HBase replication policy between two Cloudera Operational Database (COD) clusters in Cloudera Replication Manager, you can set up the triggers in Cloudera Manager to monitor specific replication-related metrics for the HBase replication policy. When the condition specified in the trigger is met, Cloudera Manager triggers an alert, and you can view the metrics for the replication policy in Cloudera Replication Manager and in Cloudera Manager.

Procedure

1. Go to the source Cloudera Manager Hosts All Hosts page.
2. Click a link in the **Name** column to open the host status page.
3. Click Create Trigger in the **Health Tests** section.

   For information about creating a trigger, see Cloudera Manager Trigger Use Cases. For information about the list of supported HBase RegionServer replication peer metrics, see HBase RegionServer Replication Peer Metrics.

4. Enter the required values and the following expression to create a trigger for a metric with a certain threshold value on the **Create New Trigger** page:

IF (SELECT *[***ENTER THE METRIC NAME***]* WHERE serviceName=$SERVICENAME AND max(*[***ENTER THE METRIC NAME***]*) > *[***ENTER THE THRESHOLD VALUE***]* AND roleType=RegionServer) DO health:concerning

The following examples illustrate how you can use triggers to monitor the replication metrics in Cloudera Manager:

a) The `IF (SELECT log_queue_size WHERE serviceName=$SERVICENAME AND max(log_queue_size) > 10 AND roleType=RegionServer) DO health:concerning` trigger statement triggers the health alert when the value of the log_queue_size is greater than 2.

This trigger condition is met when the processing of the queue becomes slow which results in the logs being queued or the replication process being halted.

b) The following sample image shows the chart that appears on the source Cloudera Manager Charts page when the health alert is triggered for the following trigger expression: `IF (SELECT shipped_batches_rate WHERE serviceName=$SERVICENAME AND max(shipped_batches_rate) > 0.01 AND roleType=RegionServer) DO health:concerning`
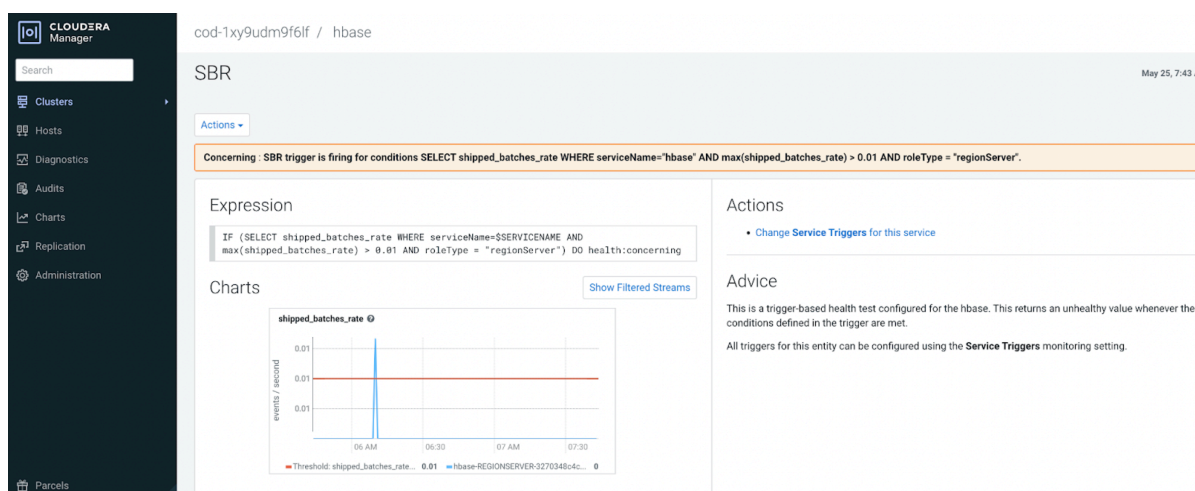
c) The following image shows the **Chart Builder** where you can configure the options as required and save it for troubleshooting purposes or for future references.
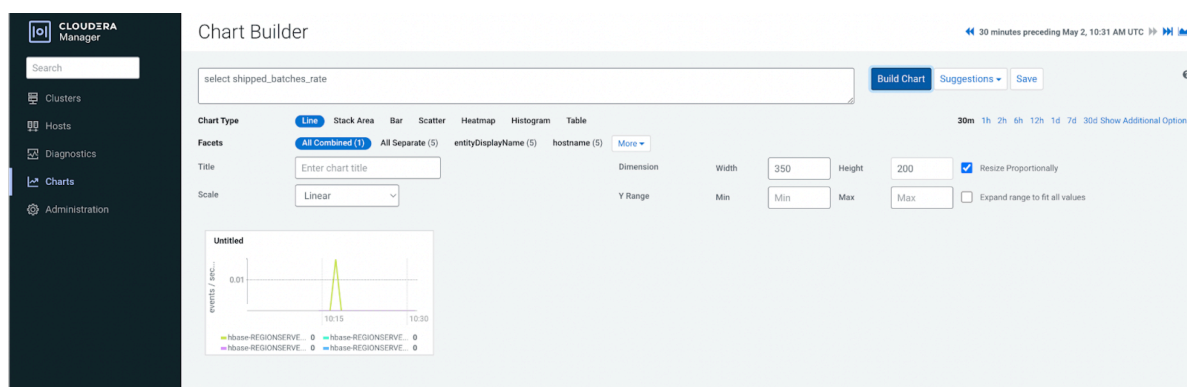
For information about monitoring the metrics related to replication and HBase health on your COD environment, see Monitor COD metrics.

What to do next

You can view and monitor the metrics in Cloudera Manager and on the **Charts** page in Cloudera Replication Manager.

Monitor HBase RegionServer replication peer metrics in Replication Manager

When you click a replication policy on the "Replication Policies" page, the "Charts" tab appears. On this tab, when you choose the HBase replication peer metric, aggregate rollup level, start time, and end time, a graphical representation of HBase RegionServer metrics specific to the replication peer appears.

Cloudera Manager monitors several metrics which includes performance metrics for the HBase RegionServer. Replication Manager leverages this capability to display the HBase RegionServer metrics specific to a replication peer as a graph on the **Charts** tab for a replication policy, and stores it in the source cluster of the replication policy. You can use these metrics to monitor a HBase replication job and to find and diagnose issues related to the HBase replication peer.

You can view these metrics for a replication peer only if the following conditions are true:

- The source Cloudera Manager API version is 42 or higher and the target Cloudera Manager API version is 53 or higher.

    Tip: The endpoint http://*[***CM HOST***]:[***CM PORT***]*/api/version shows the API version of the Cloudera Manager.

- The HBase replication policy has been successfully created and is not in an erroneous state.

    Note: If the target Cloudera Manager version is 45 or higher, the graphs generated for HBase replication policies are the same if they share the same source and target clusters. If the target Cloudera Manager version is lower than 45, the graph differs from one HBase replication policy to another because one replication peer is created for each HBase replication policy in these API versions.

For more information about the available metrics, see HBase RegionServer Replication Peer Metrics.

Viewing HBase RegionServer replication peer metrics

You can view HBase RegionServer metrics specific to a replication peer as a graph on the "Charts" tab for a replication policy on the "Replication Policies" page.

Procedure

1. Go to the **Replication Policies** page in Cloudera Replication Manager.
2. Click a successful HBase replication policy that is not in an erroneous state.
3. Go to the **Charts** tab.

4. Choose the following options depending on your requirement:

   a) Choose one of the following HBase replication peer metric:

      - Age of last shipped operation
      - Age of last shipped operation rate
      - Log edits filtered rate
      - Log edits read rate
      - Log queue size
      - Log read in bytes rate
      - Shipped batches rate
      - Shipped ops rate
      - Shipped size in kb rate
      - Shipped hfiles rate
      - Size of hfile refs queue

   b) Choose an aggregate rollup level for the metrics:

      - Raw
      - 10 minutes
      - 1 hour
      - 6 hours
      - Daily
      - Weekly

   c) Choose a Start Time.

   d) Choose an End Time.

5. Click Load Metrics to view the graphs.

Troubleshooting replication policies in Cloudera Replication Manager

The troubleshooting scenarios in this topic help you to troubleshoot issues in the Cloudera Replication Manager.

Different methods to identify errors related to failed replication policy

What are the different methods to identify errors while troubleshooting a failed replication policy?

Procedure

You can choose one of the following methods to identify the errors to troubleshoot a job failure:

- On the **Replication Policies** page, click the failed job in the **Job History** pane. The errors for the failed job appear.

  The following sample image shows the **Job History** pane for a replication policy job:

- In the source and target Cloudera Manager, click **Running Commands** on the left navigation bar. The recent command history shows the failed commands.

  The following sample image shows the **Running Commands** page for an HBase replication policy:

- On the source cluster and target cluster, open the service logs to track the errors (For example, HBase service logs).

  You can also search on the Cloudera Manager Diagnostics Logs page to view the logs.

Replication Policies page does not display all the replication policies

The "Replication Policies" page might not display all the replication policies depending on various factors. In such scenarios, you can choose to reload the page, choose a load page option, or use CDP CLIs to view and monitor the replication policies and its statistics.

Problem

When a Cloudera Manager instance is slow, that is while handling more than 650 replication policies or when it is generally under heavy load, it might slow down the 'policy list request' operation. In such scenarios, the replication policies take more time than expected to appear, or might not get displayed on the **Replication Policies** page.

Solution

Procedure

- Force reload page using the ⟳ option on the **Replication Policies** page.

  You can use this option if the 'policy list request' operation has timed out on the Cloudera Manager. You can identify this scenario when the cluster that stores the replication policies shows up in the Error list. Click

  ⚠

  to see the error list and the list of unreachable clusters.

  > **Tip:** Sometimes, Replication Manager fails to reach a healthy Cloudera Manager when there is a temporary networking blip or when there is a load spike on Cloudera Manager. When a cluster becomes unreachable for Replication Manager, the cluster is placed in the list of unreachable clusters. Replication Manager retries to reach the cluster again after 20 minutes. After you confirm that the Cloudera Manager is healthy and expect it to be reachable by Replication Manager, you can force reload the **Replication Policies** page to reconnect every cluster.

- Omit the job history of the policies to speed up the 'policy list query' operation using the Never load history option.

  Tip: Depending on your requirements, you can choose one of the following options in

  to load the **Replication Policies** page faster by delaying to load the job history:

  - Delay loading job history when it takes too long attempts to load the job history, but omits the load operation above a certain threshold. By default, Replication Manager uses this option.
  - Never load job history minimizes the load on Cloudera Manager and maximizes Replication Manager performance.
  - Always load job history ensures that the job history is always loaded for all the displayed replication policies.

  The following sample image shows the options for

  :
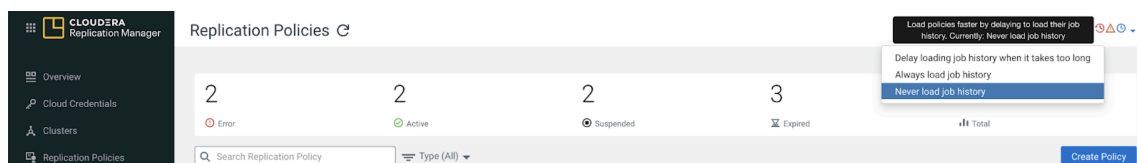
- If the Replication Manager UI still performs slowly and you are not able to view the job history and monitor the dashboards, Cloudera recommends that you use CDP CLIs for Replication Manager. The CDP CLI commands for Replication Manager are under the replicationmanager CDP CLI option.

  For more information, see CDP CLI for Replication Manager.

  Important: Policy statistics and job details are incorrect if the job history for a replication policy does not load completely. This is because the job history is necessary to decide whether a policy failed or succeeded.

  You can identify this scenario when warnings appear on the **Overview** or **Replication Policies** pages. The following sample images shows the warnings that appear on the **Overview** and **Replication Policies** pages respectively:
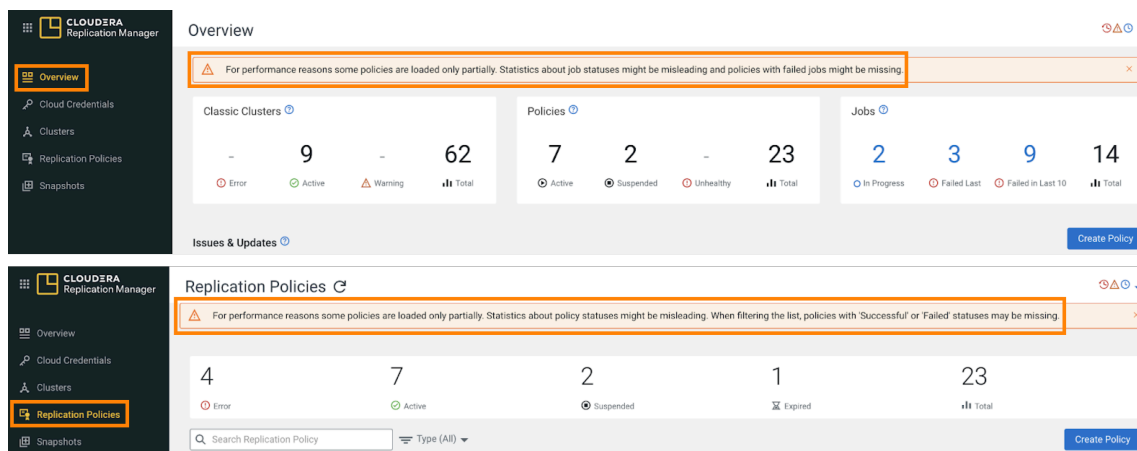
HDFS replication policy fails due to export HTTPS_PROXY environment variable

HDFS replication policies fail when the export HTTPS_PROXY environment variable is set to access AWS through proxy servers. How to resolve this issue?

Remedy

Procedure

To resolve this issue, perform the following steps:

1. Open the core-site.xml file on the source cluster.
2. Enter the following properties in the file:

```
<property>
  <name>fs.s3a.proxy.host</name>
  <description>Hostname of the (optional) proxy server for S3 connection
s.</description>
</property>

<property>
  <name>fs.s3a.proxy.port</name>
  <description>Proxy server port. If this property is not set
    but fs.s3a.proxy.host is, port 80 or 443 is assumed (consistent with
    the value of fs.s3a.connection.ssl.enabled).</description>
</property>
```

3. Save and close the file.
4. Restart the source Cloudera Manager.
5. Run the failed HDFS replication policies in Replication Manager.

    Replication Manager completes the replication successfully.

Cannot find destination clusters for HBase replication policies

When you ping destination clusters using their host names, the source cluster hosts for HBase replication policies do not find the destination clusters. How to resolve this issue?

Cause

This might occur for on-premises clusters such as Cloudera Base on premises clusters or CDH clusters because the source clusters are not on the same network as the destination Data Hub. Therefore, hostnames cannot be resolved by the DNS service on the source cluster.

Remedy

Procedure

Add the destination Region Server and Zookeeper IP to host name mappings in the /etc/hosts files of all the Region Servers on the source cluster.

The following snippet shows the contents in a sample /etc/hosts file:

```
10.115.74.181 dx-7548-worker2.dx-hbas.x2-8y.dev.dr.work
10.115.72.28 dx-7548-worker1.dx-hbas.x2-8y.dev.dr.work
10.115.73.231 dx-7548-worker0.dx-hbas.x2-8y.dev.dr.work
10.115.72.20 dx-7548-master1.dx-hbas.x2-8y.dev.dr.work
10.115.74.156 dx-7548-master0.dx-hbas.x2-8y.dev.dr.work
10.115.72.70 dx-7548-leader0.dx-hbas.x2-8y.dev.dr.work
```

HBase replication policy fails when Perform Initial Snapshot is chosen

An HBase replication policy fails for COD on Microsoft Azure when the "Perform Initial Snapshot" option is chosen but data replication is successful when the option is not chosen. How to resolve this issue?

Cause

This issue appears when the required managed identity of source roles are not assigned.

Remedy

Procedure

Assign the managed identity of source roles, Storage Blob Data Owner or Storage Blob Data Contributor, to the destination storage data container and vice versa for bidirectional replication.

The roles allow writing a snapshot in the destination cluster container.

Optimize HBase replication policy performance when replicating HBase tables with several TB data

Can HBase replication policy performance be optimized when replicating HBase tables with several TB of data if the "Perform Initial Snapshot" option is chosen during HBase replication policy creation?

Complete the following manual steps to optimize HBase replication policy performance when replicating several TB of HBase data if you choose the Perform Initial Snapshot option during the HBase replication policy creation process.

Remedy

Procedure

1. Before you create the HBase replication policy, perform the following steps:
    a) Navigate to the source Cloudera Manager YARN service Configuration tab.
    b) Search for the mapreduce.task.timeout parameter.
    c) Increase the value or set it to 0 to switch off the timeout.
    d) Restart the YARN service.
    e) Navigate to the source Cloudera Manager HBase service Configuration tab.
    f) Search and configure the following key-value pairs:

    - hbase.snapshot.master.timeout.millis = 840000
    - hbase.client.sync.wait.timeout.msec = 180000
    - hbase.client.operation.timeout = 2400000
    - hbase.client.procedure.future.get.timeout.msec = 3000000
    - hbase.hfilearchiver.thread.pool.max=100
    - hbase.snapshot.thread.pool.max=24
    g) Restart the HBase service.
    h) Perform steps e through g on the target Cloudera Manager.
2. When you create the HBase replication policy for the first time using the above configured source cluster, you must increase the Maximum Map Slots value to a higher number on the Advanced Settings page.

    Tip: Calculate this number by multiplying the number of nodes on the source cluster and the number of cores in the node.

3. If Store File Tracking (SFT) is enabled in the target COD, perform the steps mentioned in the COD migration topic after the replication policy creation is complete.

    Note: SFT is enabled by default on Cloudera clusters with COD version 7.2.14.2 and higher using Cloudera Manager versions 7.2.16 and higher.

Partition metadata replication takes a long time to complete

How can partition metadata replication be improved when the Hive tables use several Hive partitions?

Hive metadata replication process takes a long time to complete when the Hive tables use several Hive partitions. This is because the Hive partition parameters are compared during the import stage of the partition metadata replication process and if the exported and existing partition parameters do not match, the partition is dropped and recreated. You can configure a key-value pair to support partition metadata replication.

Procedure

1. Go to the Cloudera Manager Clusters *HIVE SERVICE* Configuration tab.

2. Search for the Hive Replication Environment Advanced Configuration Snippet (Safety Valve) property.

3. Enter the HIVE_IGNORED_PARTITION_PARAMETERS=*[***COMMA SEPARATED LIST OF HIVE PARTITION PARAMETERS***]* key-value pair.

   For example,

   ```
   HIVE_IGNORED_PARTITION_PARAMETERS=transient_lastDdlTime,totalSize,numRow
   s,COLUMN_STATS_ACCURATE,numFiles,numFilesErasureCoded
   ```

   The partition parameter names you provide are not compared during the import stage of the partition metadata replication process. Therefore, even if the partition parameters do not match between the exported and existing partitions, the partition is not dropped or recreated. After you configure this key-value pair, the import stage of the partition metadata replication process completes faster.

4. Save the changes, and restart the Hive service.

Replicating Hive nested tables

Cloudera Replication Manager does not support Hive nested tables. What do I do if there are Hive nested tables in the source cluster?

Cloudera Replication Manager does not support Hive nested tables for replication. Therefore, it is recommended that you move the nested tables to a different location in HDFS and then replicate Hive external tables. However, if this is not possible, you can perform the following steps in the given order as a workaround.

Solution

Procedure

1. Create a Hive replication policy on the target cluster. Ensure that the Additional Settings Replication Option Metadata only option is selected to replicate the metadata of required files and directories.

   Note: If you are using Cloudera on cloud 7.2.15 or lower and if you are using Amazon S3 as the disaster recovery cluster, ensure that you (the administrator) run the following command after the Hive metadata replication is complete:

   ALTER TABLE table SET LOCATION "*[***[S3A://S3_BUCKET/WAREHOUSE_PATH]***]*";

2. Create a HDFS replication policy on the source cluster to replicate the table data.

Target HBase folder is deleted when HBase replication policy fails

When the snapshot export fails during the HBase replication policy job run, the target HBase folder in the destination Data Hub or COD gets deleted.

Note: This scenario appears if you are using Cloudera Manager versions that are lower than 7.6.7 CHF8, 7.11.0, or 7.9.0-h6 on the source cluster.

You can either revoke the delete permission for the user, or ensure that you use an access key/role that does not have delete permissions to the required storage component.

The following steps show how to create an access key in AWS and an Azure service principal, which do not have delete permission for the storage component.

Solution in AWS

Procedure

If the Data Hub or COD is on Amazon S3, you can perform the following steps to create an access key that does not have delete permission for the storage component:

1. Login to AWS.
2. Click Create policy on the IAM Policies Create policy page.

   The following sample image shows the **Policies** page in the AWS IAM section to create a policy.
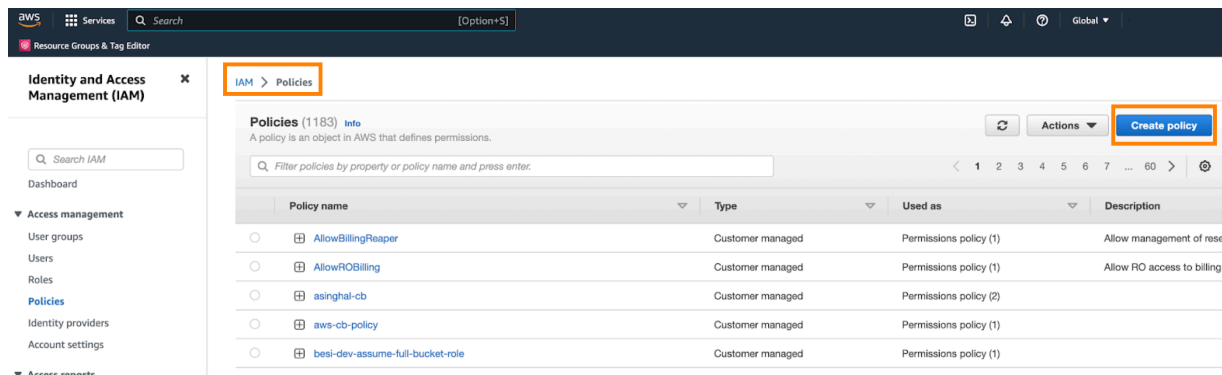
3. Choose the S3 Service, and then choose the following permissions for generic access (assign resources as required):

   - *List/ListBucket*
   - *Read/GetBucketLocation*
   - *Read/GetObject*
   - *Write/AbortMultipartUpload*
   - *Write/PutObject*
   - *Permissions management/PutBucketPublicAccessBlock*

4. Add *Delete/DeleteObject* permission to the target COD cluster's snapshot temporary folder.
   For example, the target COD cluster's snapshot temporary folder might be located in *[***TARGET COD S3 PATH***]*/hbase/.hbase-snapshot/.tmp/*.

5. Enter a Name for the policy, add tags, and click Create policy.

6. Click Add Users to create a user on the IAM Users page.

7. Enter a Name, and click Next.

8. Choose the Attach policies directly option on the **Set permission** page, and then assign the previously created policy to the user.

9. Optionally, add tags and create the user.

10. Click Create access key to create an access key for the user on the IAM Users *[***NEW USER***]* page.

11. On the Security Credentials Access keys page, choose Application running outside AWS. Click Next.

12. Optionally, attach the tags, create and save the access key. This access key is used as an external account for replication.

How do I verify whether the target HBase folder in the destination Data Hub or COD does not get deleted if the snapshot export fails during the HBase replication policy job run?

Perform the following steps to verify if the delete operation is allowed for the access key that you previously created:

a. Run the aws configure --profile delete-test command to setup the credentials in AWS CLI.

b. Delete an arbitrarily created temporary file from the account using the aws s3 --profile delete-test rm -- recursive s3://*[***ACCOUNT NAME***]*/delete-testing/ command.

The delete operation is not allowed.

Solution in Microsoft Azure

Procedure

If the Data Hub or COD is on ABFS, you can perform the following steps to create an Azure service principal that does not have delete permission for the storage component:

1. Login to Microsoft Azure.

2. Click Add Add a custom role on the Subscriptions Select the subscription Access Control (IAM) page in Microsoft Azure, and complete the following steps:

    a) On the **Basics** tab, provide a name for the role, select Clone a role for the **Baseline permissions** field, and choose Storage Blob Data Contributor for the **Role to clone** field. Click Next.

    b) On the **Permissions** tab, remove the *Microsoft.Storage/storageAccounts/blobServices/containers/delete* and *Microsoft.Storage/storageAccounts/blobServices/containers/blobs/delete* permissions.

    c) Click Review + create.

The following sample image shows the Access control (IAM) page in Microsoft Azure:
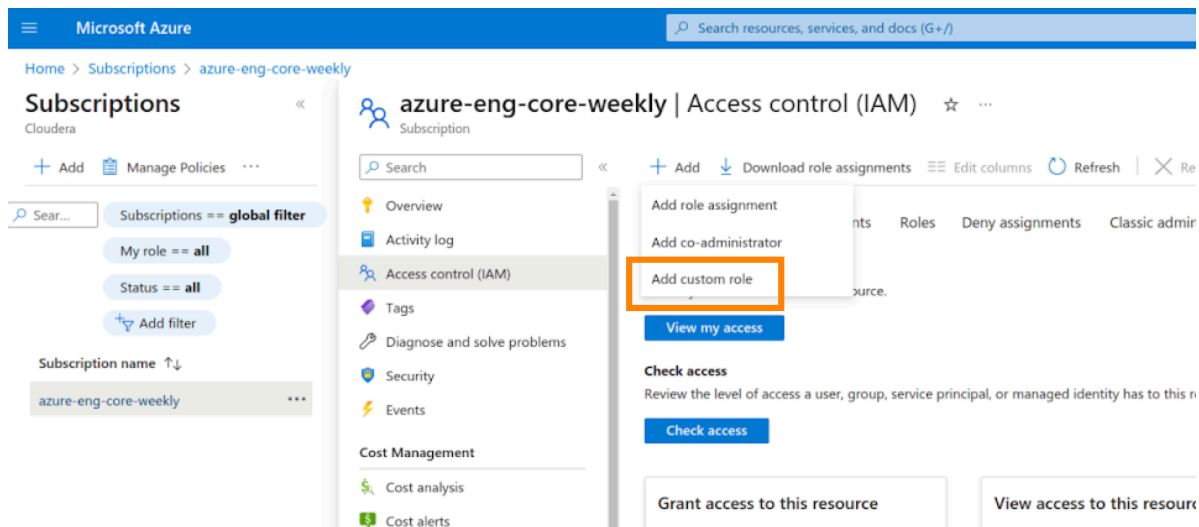
3. Click Add Add role assignment on the Storage accounts Select the account where snapshots will be written Access Control (IAM) page, and complete the following steps:

    a) On the **Role** tab, select the custom role previously created. Click Next.

    b) On the **Members** tab, select User, group, or service principal for **Assign access to** field, and select the required service principal.

    c) Click Review + assign.

    d) Click Review + assign on the **Conditions (optional)** tab.

4. Click Add principal on the Storage accounts *[***ACCOUNT NAME***] [***CONTAINER WHERE SNAPSHOTS ARE TO BE WRITTEN***]* Manage ACL page.

    a) Select the required service principal.

    b) Choose the Execute permission for the required container, and click Save.

> ⚠️ **Important:** Ensure that no other role containing delete access to the given storage account is assigned to the new service principal on the Access Control (IAM) page.

How do I verify whether the target HBase folder in the destination Data Hub or COD does not get deleted if the snapshot export fails during the HBase replication policy job run?

To verify if the delete operation is allowed on the service principal that you previously created, perform the following steps:

a. Open the Azure Cloud Shell terminal.

b. Login using the service principal that you created previously using the az login --service-principal -u *[***CLIENT ID***]* -p *[***CLIENT SECRET***]* --tenant *[***TENANT ID***]* command.

c. Delete an arbitrarily created temporary file from the account using the az storage fs file delete --path *[***TEMPORARY FILE***]* -f data --account-name *[***ACCOUNT NAME***]* --auth-mode login command.

The delete operation is not allowed.

Replicate HBase data in existing and future tables

Errors might appear when you try to replicate HBase data from existing tables and future tables in a database using the "Replicate Database" option during the HBase replication policy creation process. These errors appear when there are compatibility issues.

The following list shows a few errors that might appear and how to mitigate these issues:

Procedure

- The error *Destination cluster doesn't support replication of all HBase tables. Please change the destination cluster or specify the tables that should be replicated.* appears during HBase replication policy creation process. How to mitigate this issue?

  This error appears if you choose the Select Source Replicate Database option during the HBase replication policy creation process for unsupported cluster versions.

  To mitigate this issue, ensure that the target Cloudera Manager version is 7.11.0 or higher and the source cluster version is CDH 6.x or higher before you choose the Select Source Replicate Database option during the HBase replication policy creation process.

- The error *Exception in thread "main" java.lang.IllegalArgumentException: Cannot add a peer with id=_repl__f1907207cdbc4a528777bb4a316_ba23f09f7328494bbcbf81f40b because that id already exists* appears after creating subsequent HBase replication policies.

  This error appears if you created an HBase peer manually using the *hbase shell* to replicate all column families on a source cluster using CDH 5.x.

  You can replicate the HBase data (existing tables and future tables) in a database only if the target Cloudera Manager version is 7.11.0 or higher and the source cluster version is CDH 6.x or higher. Therefore, even if you create a peer using *hbase shell* in the source cluster version CDH 5.16.2 or any other unsupported version, errors might appear.

Appendix

Before you create replication policies, you must register the Amazon S3 or Azure cloud credentials to use as cloud storage in Cloudera Replication Manager, and register the on-premises clusters (CDH or Cloudera Base on premises) as classic clusters in the Management Console.

Support matrix for Cloudera Replication Manager

You can use Replication Manager or other alternate replication methods to replicate HDFS, Hive external tables, and HBase data between on-premises clusters (CDH clusters, Cloudera Base on premises clusters, HDP clusters) and Cloudera on cloud (Amazon S3 (AWS), Microsoft Azure ADLS Gen2 (ABFS), and Google Cloud Platform (GCP)) clusters. Replication Manager from HDP clusters to Cloudera Public Cloud Azure is a beta feature and is not available for general use.

> **Note:** Before you create replication policies, you must verify whether the on-premises cluster versions are supported by Replication Manager, register the on-premises clusters as classic clusters in the Cloudera Management Console, register the cloud account credentials in the Replication Manager service, verify cluster access, and configure minimum ports for replication.
>
> Ensure that the target cluster's and source cluster's and versions are the same, or the target cluster's version is higher than the source cluster's version.

See the other sections in this topic for the lowest supported cluster and runtime versions:

- Replication policies and features.
- Cloudera Base on premises and Cloudera on cloud source clusters that Replication Manager supports.
- CDH and HDP source clusters that Replication Manager supports.

List of features supported by Cloudera Replication Manager

Replication Manager provides replication policies that you can create, edit, and manage to accomplish your data replication goals. You can use other alternate replication methods for scenarios that Replication Manager does not support. Certain features in Cloudera Replication Manager are available only if the source and target clusters' Cloudera Manager versions support the feature. Verify whether your source and target cluster's Cloudera Manager version support the required feature.

Supported replication policies

You can use the following replication policies in Cloudera Replication Manager:

- HDFS replication policies
- Hive replication policies
- HBase replication policies

HDFS replication policies

Replicate HDFS data and metadata from:

- on-premises clusters (CDH, Cloudera Base on premises, and HDP) to cloud storage.
- cloud storage to classic clusters (CDH or Cloudera Base on premises clusters).

You can choose the frequency during policy creation to replicate the data.

Hive replication policies

Support table-level replication, and can replicate Hive external tables from on-premises clusters (CDH and Cloudera Base on premises) to cloud storage and to Data Hubs. The replication policies can also:

- replicate data stored in Hive tables, Hive metadata, data in Hive metastore, and Impala metadata (catalog server metadata) associated with Impala tables registered in the Hive metastore, and

  > **Note:** Hive2 managed tables are converted to external tables after replication.

- migrate Sentry permissions to Ranger.

  > **Note:** To perform the Sentry policy replication, you must be running the Sentry service on CDH 5.12 or higher, or any CDH 6.x version.

You can choose the frequency during policy creation to replicate the data.

HBase replication policies

Replicate HBase data from a source classic cluster (CDH or Cloudera Base on premises cluster), COD, or Data Hub to a target Data Hub or COD cluster. You can also copy or replicate HBase data between different environments within a Virtual Private Cloud (VPC) using these policies.

Table 5: Lowest supported cluster and runtime versions for HBase replication policies

Source Cluster Type	Lowest Supported Source CDH/ Cloudera Version	Lowest Supported Source Cloudera Manager Version	Target Cluster Type	Lowest Supported Target Cloudera Version	Lowest Supported Target Cloudera Manager Version
Cloudera Base on premises	7.1.6*	7.3.1	Data Hub in Cloudera on cloud AWS/ Azure	7.2.14	7.6.0
CDH	6.3.3	7.3.1	Data Hub in Cloudera on cloud AWS/ Azure	7.2.14	7.6.0
CDH	5.16.2	7.4.4 (patch-5017)	COD (AWS)	7.2.14	-
CDH	5.16.2	• 7.6.1 (patch-5610) • 7.6.7 CHF1 and higher	COD (Azure)	7.2.14	-
COD (AWS/ Azure)	7.2.14	-	COD (AWS/ Azure)	7.2.14	-
COD (GCP)	7.2.16.1**	-	COD (GCP)	7.2.16.1**	-
COD (GCP)	7.2.16.500 7.2.17.300 7.2.18.0	-	COD (GCP)	7.2.16.500 7.2.17.300 7.2.18.0	-
Cloudera Base on premises	7.1.9 SP1	7.11.3 CHF7	GCP	7.2.16.1	-

*Cloudera Base on premises 7.1.6 and higher clusters must be Kerberos enabled to use them as source classic clusters in an HBase replication policy.

**You must add key-value pairs to register a Google account to use in Replication Manager. For more information about the key-value pairs, see Preparing to create an HBase replication policy.

⚠ **Important:**

• You can replicate HBase data in SFT-enabled clusters for target clusters with version Cloudera on cloud 7.2.16 or higher using HBase replication policies.
• You can replicate Phoenix tables from CDH 5.16.2 and higher using Cloudera Manager 7.4.4 and higher to COD using HBase replication policies.
• You cannot create HBase replication policies if the target Cloudera version is 7.2.16, 7.2.16.1, 7.2.16.2, 7.2.16.3, or 7.2.16.200, the source Cloudera Manager version is 7.7.3 or lower, and the source Cloudera Manager API version is v50 or higher.

> 💡 **Tip:** The endpoint http://*[***CM HOST***]*:*[***CM PORT***]*/api/version shows the API version of the Cloudera Manager.

• HBase replication policies do not support source or destination Cloudera on cloud clusters created using the Micro Duty template.

HBase replication policies replicate all the data from the specified tables and then continue to replicate the changed data automatically without user intervention.

> **Note:** Alternate replication methods:
> - Use the replication plugin for HBase data. For more information, see COD replication in a Nutshell, Cloudera replication plugin, and HBase data replication.
> - Contact Cloudera Support for Hive external tables.

Supported features

The following table lists the features and the lowest supported versions that are required for source clusters and target clusters to use the features:

Feature	Lowest supported source Cloudera Manager version	Lowest supported target Cloudera Manager version
Register the GCP credentials to use in Replication Manager on the **Cloud Credentials** page.	• 7.9.0-h7 and higher • 7.11.0-h3 and higher • 7.12.0.0 and higher	Supports all Cloudera on cloud Cloudera Manager versions.
Replicate HBase data simultaneously between multiple clusters*.	• 7.9.0-h7 and higher • 7.11.0-h2 and higher • 7.12.0.0 and higher	• 7.9.0-h7 and higher • 7.11.0-h2 and higher • 7.12.0.0 and higher
Replicate only those HBase tables where the replication scope is already enabled using the Select Source Replicate only tables where replication is already enabled * option during the HBase replication policy creation process.	Supports all Cloudera on cloud Cloudera Manager versions.	• 7.9.0-h7 and higher • 7.11.0-h3 and higher • 7.12.0.0 and higher
Specify the network load balancer (NLB) Endpoint after you enable the Select Destination Replicate via a Network Load Balancer* option during the HBase replication policy creation process if the on-premises cluster uses NLB to communicate with the COD clusters.	CDH 5.16.2	7.12.0.100
Specify the YARN queue bandwidth using the Initial Snapshot Settings Maximum Bandwidth * option during the HBase replication policy creation process to export the HBase initial snapshot.	• 7.9.0-h7 and higher • 7.11.0-h3 and higher • 7.12.0.0 and higher	• 7.9.0-h7 and higher • 7.11.0-h3 and higher • 7.12.0.0 and higher
Enter Initial Snapshot Settings Maximum parallel snapshots* to specify the maximum number of tables to process in parallel during the initial snapshot export and import step for an HBase replication policy. If you do not enter any value, Replication Manager chooses an appropriate value, depending on the resources in the source and target cluster, to optimize the performance.	Supports all Cloudera on cloud Cloudera Manager versions.	• 7.9.0-h7 and higher • 7.11.0-h3 and higher • 7.12.0.0 and higher
Add IDBroker credentials* to use in Replication Manager on the **Cloud Credentials** page.	7.11.3 CHF7	7.11.3 CHF7
Enter the Select Source Export snapshot user * field during the HBase replication policy creation process to specify the username to export the initial snapshot to the target.	7.11.3 CHF7	7.11.3 CHF7
*To enable this feature, contact your Cloudera Account team.		

Replicate data from Cloudera Base on premises and Cloudera on cloud source clusters

Replication Manager replicates HDFS (Cloudera Base on premises source clusters and Cloudera on cloud storage on AWS and Azure), Hive external tables (Cloudera Base on premises source clusters), and HBase (Cloudera Base on premises source clusters) data to Cloudera on cloud (Amazon S3 and Microsoft Azure ADLS Gen2 (ABFS)) clusters.

You can use the replication plugin as an alternate replication method to replicate HBase data for scenarios that are not supported by Replication Manager.

The following tables list the lowwest source and destination cluster versions, lowest Cloudera Manager versions, supported cloud providers, and supported scenarios:

Replicate data from Cloudera Base on premises source clusters

Source cluster	Lowest supported source Cloudera Manager version	Lowest supported source Cloudera Runtime version	Cloud provider	Supported services on Replication Manager	Services that require alternate replication methods
Cloudera Base on premises	7.1.1	7.1.1	Cloudera on cloud AWS/Azure	HDFS	HBase To replicate HBase data, see COD replication in a Nutshell and HBase data replication.
Cloudera Base on premises	7.1.1	7.1.1	Data Lake in Cloudera on cloud AWS/Azure	Hive external tables	
Cloudera Base on premises	7.9.0	7.1.1	Data Hub in Cloudera on cloud AWS/Azure	Hive external tables	None
Cloudera Base on premises	7.3.1	7.1.6	Data Hub in Cloudera on cloud AWS/Azure	HBase	None
Cloudera Base on premises	7.11.3 CHF7	7.1.9 SP1	Cloudera on cloud GCP	HDFS, Hive external tables, HBase	None

Note: Replication Manager converts Hive2 managed tables to external tables after replication.

Replicate data from Cloudera on cloud source clusters

Consider the following limitations while using Cloudera on cloud source and Cloudera on cloud target clusters:

- Replication across cross-cloud providers, that is from AWS to Azure and vice-versa is not supported.
- The source and target clusters must use the same account.

Table 6:

Source cluster	Lowest supported destination cluster version	Supported services on Replication Manager	Services that require alternate replication methods
Cloudera on cloud AWS* / Azure	CDH 5.x CDH 6.x HDP 2.x HDP 3.x	Not applicable	HBase To replicate HBase data, see COD replication in a Nutshell and HBase data replication.
Cloudera on cloud AWS*	CDH 5.9.0 and higher Cloudera Base on premises 7.1.7 SP1 and higher	HDFS	None
Cloudera on cloud Azure	CDH 6.1.0 and higher Cloudera Base on premises 7.1.7 SP1 and higher	HDFS	None
Cloudera on cloud GCP 7.2.18 and higher	Cloudera Base on premises 7.1.9 SP1 and higher	HDFS	None
COD version 7.2.14 and higher - Cloudera on cloud AWS	AWS	HBase	None
COD version 7.2.14 and higher - Cloudera on cloud Azure	Azure	HBase	None

Source cluster	Lowest supported destination cluster version	Supported services on Replication Manager	Services that require alternate replication methods
COD version 7.2.16.1 and higher - Cloudera on cloud GCP	GCP	HBase	None
*Replication Manager does not support S3 as a source or destination when S3 is configured to use SSE-KMS.			

Replicate data from CDH and HDP source clusters

Replication Manager replicates HDFS data (CDH source clusters and HDP source clusters), Hive external tables (CDH source clusters), and HBase data (CDH 6 source clusters) to Cloudera on cloud (Amazon S3 and Microsoft Azure ADLS Gen2 (ABFS)) clusters. Replication Manager from HDP clusters to Cloudera on cloud Azure is a beta feature and is not available for general use. You can use alternate methods to replicate Hive external tables and HBase data for scenarios that are not supported by Replication Manager.

The following tables list the lowest CDH and HDP source cluster versions, lowest Cloudera Manager versions, supported cloud providers, and supported scenarios:

Table 7: Replicate data from CDH 5 source clusters

Source cluster	Lowest supported source Cloudera Runtime version	Lowest supported source Cloudera Manager version	Cloud provider	Supported services on Replication Manager	Services that require alternate replication methods
CDH 5	5.10	6.3.0	Cloud storage in Cloudera on cloud AWS	HDFS	HBase To replicate HBase data, see COD replication in a Nutshell, Migrating HBase data, and HBase data replication.
CDH 5	5.10	6.3.0	Data Lake in Cloudera on cloud AWS	• Sentry to Ranger* • Hive external tables	
CDH 5	5.10	6.3.4	Cloud storage in Cloudera on cloud Azure	HDFS	
CDH 5	5.10	6.3.4	Data Lake in Cloudera on cloud Azure	• Sentry to Ranger* • Hive external tables	
CDH 5	5.10	7.9.0	Data Hub in Cloudera on cloud AWS/Azure	Hive external tables	None
*To perform the Sentry policy replication, you must be running the Sentry service on CDH 5.12 or higher, or any CDH 6.x version.					

Table 8: Replicate data from CDH 6 source clusters

Source cluster	Lowest supported source Cloudera Runtime version	Lowest supported source Cloudera Manager version	Cloud provider	Supported services on Replication Manager	Services that require alternate replication methods
CDH 6	6.1	6.3.0	Cloud storage in Cloudera on cloud AWS	HDFS	HBase To replicate HBase data, see COD replication in a Nutshell, Migrating HBase data, and HBase data replication.
CDH 6	6.1	6.3.0	Data Lake in Cloudera on cloud AWS	• Sentry to Ranger* • Hive external tables	
CDH 6	6.1	7.1.1 / 6.3.4	Cloud storage in Cloudera on cloud Azure	HDFS	

Source cluster	Lowest supported source Cloudera Runtime version	Lowest supported source Cloudera Manager version	Cloud provider	Supported services on Replication Manager	Services that require alternate replication methods
CDH 6	6.1	7.1.1 / 6.3.4	Data Lake in Cloudera on cloud Azure	• Sentry to Ranger* • Hive external tables	
CDH 6	6.1	7.9.0	Data Hub in Cloudera on cloud AWS/Azure	• Sentry to Ranger* • Hive external tables	
CDH 6	6.3.3	7.3.1	Data Hub in Cloudera on cloud AWS/Azure	HBase	None

*To perform the Sentry policy replication, you must be running the Sentry service on CDH 5.12 or higher, or any CDH 6.x version.

> **Note:** Replication from HDP clusters to Cloudera on cloud clusters is a technical preview feature and is not available for general use. Before you replicate data from HDP, contact your Cloudera account team.

Table 9: Replicate data from HDP 2 and HDP 3 source clusters

Lowest supported source HDP version	Cloud provider	Supported services on Replication Manager	Services that require alternate replication methods
HDP 2.6.5*	AWS	HDFS	• HBase To replicate HBase data, see COD replication in a Nutshell and HBase data replication. • Hive external tables For more information, contact Cloudera Support.
HDP 2.6.5*	Azure	HDFS	HBase To replicate HBase data, see COD replication in a Nutshell and HBase data replication.
HDP 3.1.1*	AWS Azure	HDFS	• HBase To replicate HBase data, see COD replication in a Nutshell and HBase data replication. • Hive external tables For more information, contact Cloudera Support.

*No alternate replication methods are available for HDFS, Ranger, and Atlas replication.

Cloud credentials to use in Cloudera Replication Manager

The Cloud Credentials page shows the registered cloud credentials for Replication Manager. To replicate data to a storage cloud account, you must register the cloud credentials, so that the Replication Manager can access your cloud account. The supported cloud storage accounts are Amazon S3 and Azure Blob Filesystem (ABFS). On the Cloud Credentials page, you can add, update, or delete cloud credentials. Before you register an Amazon S3 cloud account, ensure the cloud bucket requirements are met. Before you add Azure Cloud Credentials in Cloudera Manager, ensure the Blob container requirements are met.

You can perform the following tasks on the Cloud Credentials page to manage cloud credentials:

Add cloud credentials

You can add cloud credentials for your S3 or ABFS account. For information about adding cloud credentials, see Working with Cloud Credentials.

Note: Unregistered credentials can impact the replication process. Credentials associated with a cluster node that do not have updated credentials are called unregistered credentials. For example, if a node is down when the credentials are changed on a bucket or when the node is brought up that has the old credentials.

Update cloud credentials

You can update the cloud credentials based on various factors. When the bucket configuration such as secret or access keys, bucket name or endpoint, and encryption type is changed, it can affect the Replication Manager replication policy run and might require an update to the Replication Manager cloud credentials.

Credential changes are picked up by the next run of the policy. When you change the credentials, the in-progress policy runs might fail but the succeeding runs pick up the changes.

To update a cloud credential, click Actions Update option.

Delete cloud credentials

You can delete unwanted credentials from the Replication Manager. When you delete cloud credentials, the replication policies that use the deleted cloud credentials might fail. To avoid failures, delete the Replication Manager cloud policies associated with the deleted credentials and recreate the policies with the new credentials. You can view a list of policies associated with specific credentials on the **Cloud Credentials** page.

To delete a cloud credential, click Actions Delete option.

Registering Amazon S3 cloud account in Replication Manager

You must have valid Amazon S3 credentials to register the cloud account with Replication Manager.

Before you begin

Consider the following requirements before you register an Amazon S3 cloud account in Replication Manager:

- You need a cloud bucket with user credentials that you can enter in Replication Manager, so Replication Manager can access the bucket.
- The bucket has to have enough space for the replicated data, and write permissions to copy the data.
- The bucket needs to support cloud storage encryption types supported by Replication Manager (SSE-S3 & SSE-KMS).

About this task

When you add cloud credentials for your Amazon S3 account, you can choose one of the following authentication methods:

- Access secret key. To use this authentication type, you require an AWS Access Key and an AWS Secret key that you obtain from Amazon. Cloudera Manager stores these values securely and does not store them in world-readable locations. The credentials are masked and encrypted in the configurations passed to processes managed by Cloudera Manager, and redacted from the logs.

- IAM role. Amazon Identity and Access Management (IAM) can be used to create users, groups, and roles for use with Amazon Web Services, such as EC2 and Amazon S3. IAM role-based access provides the same level of access to all clients that use the role.

  ⚠️ **Important:** IAM role conditions

  You can choose the IAM role authentication type only when the following conditions are met:

  - The source cluster is hosted on an AWS EC2 infrastructure.
  - The source cluster Cloudera Manager and all the nodes in the cluster are running on an EC2 instance.
  - The source cluster Cloudera Manager has the same IAM role.

  For information about configuring AWS credentials, see Introduction to role based provisioning credential in AWS.

Procedure

1. Go to Replication Manager Cloud Credentials page, and click Add.
2. In the Add Cloud Credential window, perform the following steps:
   a) Select the Cluster.
   b) Select S3 as the Cloud Storage Type.
   c) Name - Provide a unique cloud credential name.
   d) Authentication Type - Select one of the following authentication type:
      - Select the authentication type as Access Secret Key from the drop-down.
        - Access Key - Enter the valid access key.
        - Secret Key - Enter the valid secret key.
      - Select IAM Role if the conditions mentioned in the IAM Role conditions section are met, and click Save.

3. Click Validate.

> **Note:** Using the validation feature is recommended to ensure that the Amazon S3 bucket keys are valid. If the keys are not valid, the Replication Manager policy cannot execute a copy of data to the target Amazon S3 bucket.

Add Cloud Credential

Cluster

Select...

Cluster 1 (powqwehugt) ✕

Cloud Storage Type

S3

Name *

Enter a unique name for the cloud credential

Authentication Type

Access & Secret Key

Access Key *

Enter S3 access key

Secret Key *

Enter S3 secret key

Cancel   Validate

What to do next

Verify whether the credentials are listed on the Cloud Credentials page.

Register Azure cloud credentials in Replication Manager

You require an ADLS Gen2 storage account that has a cloud Blob container with user credentials to use in the Replication Manager service. The container must have enough space for the replicated data, and write permissions to copy the data.

Currently, registering Azure cloud credentials using the Replication Manager UI does not automatically create the same authorisation rules on the source Cloudera Manager cluster. When you plan to submit the replication policies with Azure as your cloud storage, Cloudera recommends that you update the cloud credentials in the source Cloudera Manager.

Registering ABFS cloud account in Replication Manager

You must have valid ABFS credentials in Cloudera Replication Manager to register the cloud account.

Before you begin

Consider the following requirements before you register an ABFS cloud account in Replication Manager:

- You require an ADLS Gen2 storage account that has a cloud Blob container with user credentials to use in the Replication Manager service.
- The container must have enough space for the replicated data, and write permissions to copy the data.
- The container must support cloud storage encryption types supported by Replication Manager (SSE-S3 & SSE-KMS).

Procedure

1. Click Cloudera Replication Manager Cloud Credentials Add .
2. Perform the following steps on the Add Cloud Credential modal window:
   a) Select the Cluster.
   b) Select ABFS as the Cloud Storage Type.
   c) Enter the cloud credential Name.
   d) Enter your ABFS Storage Client Id.
   e) Enter your ABFS Storage Client Secret Key.
   f) Enter your ABFS Storage Tenant Id.
3. Click Validate.

After you add the ABFS cloud credentials and you create a replication policy with ABFS as your selected cloud storage for your target cluster, the following error message might appear:

Error   ✕

java.lang.RuntimeException: com.cloudera.cmf.service.config.ConfigGenException: Required account config value not found: adls_tenant_id

OK

To resolve this issue, update the ABFS cloud credential values in the source Cloudera Manager instance. For more information, see *Updating Azure Cloud Credentials in Cloudera Manager*.

What to do next
Verify whether the credentials appear on the **Cloud Credentials** page.
Updating Azure Cloud Credentials in Cloudera Manager
Before you register an ABFS cloud account in Replication Manager, Cloudera recommends that you update the cloud credentials in the source cluster Cloudera Manager. This is because registering Azure cloud credentials using the Replication Manager UI does not automatically create the same authorisation rules on the source Cloudera Manager cluster.

Procedure

1. Click Add AD Service Principal on the Cloudera Manager Administration Azure Credentials page for the source cluster instance.
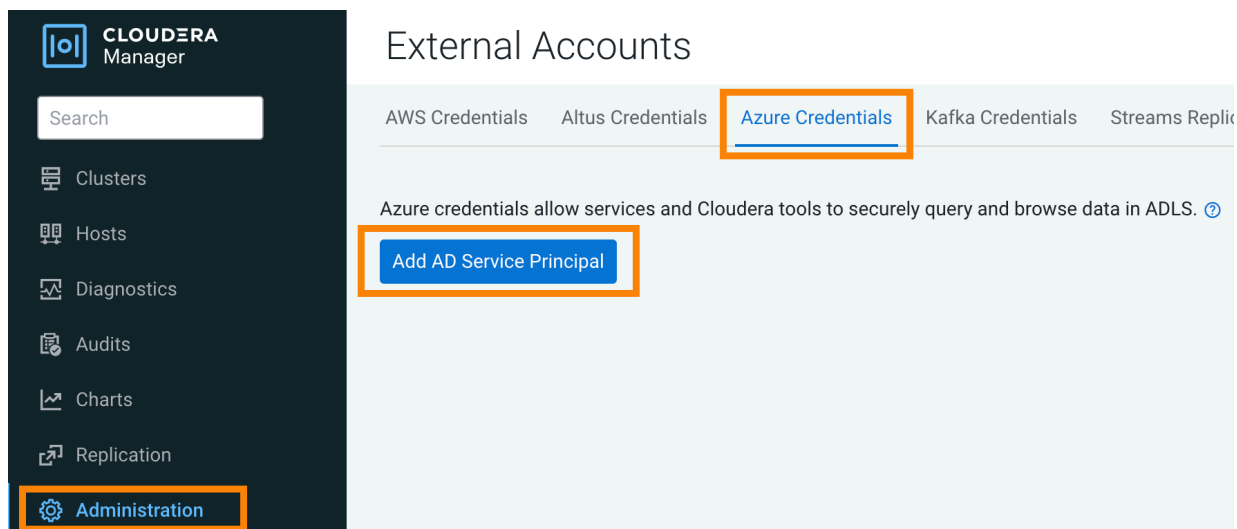
2. Enter the following details on the **Add AD Service Principal** modal window:
   a) Name of the credential.
   b) Client ID for the credential.
   c) Client Secret Key for the credential.
   d) Tenant ID for the credential.

3. Click Add.
The Azure Cloud credentials are successfully registered with the Cloudera Manager instance.

What to do next
Register the ABFS cloud account in Cloudera Replication Manager.

Registering GCP credentials to use in Replication Manager
You can register Google Cloud Storage (GCS) credentials to use in Replication Manager after you provide access to GCP for the source cluster in the source Cloudera Manager.

Before you begin
Ensure that the following details are available before you register the GCS credentials to use in Replication Manager.

- A valid Google Cloud Platform service account. For more information, see Configuring access to Google Cloud Storage.
- Enough space for the replicated data and the required write permissions to copy the data in the storage of the destination cluster.

Procedure

1. Click Cloud Credentials Add inCloudera Replication Manager.

2. Select the required source Cluster.

3. Choose GCS as the Cloud Storage Type.

4. Enter a unique Name for the cloud credential.

5. Select one of the following Input Type, and then complete the required steps:

    a) Select Manual if you have the following details about the GCP service account:

    - Enter your GCS service account Client Email address.
    - Enter your GCS service account Private Key.
    - Enter your GCS service account Private Key ID.

Add Cloud Credential   ✕

Cluster

Select...   ⌄

Cloud Storage Type

GCS   ⌄

Name *

Enter a unique name for the cloud credential

Input Type

Manual   ⌄

Client E-mail *

Enter your GCS Service Account Client E-mail

Private Key *

Enter your GCS Service Account Private Key

Private Key ID *

Enter your GCS Service Account Private Key ID

Cancel   Save

b) Select File if you chose to save the service account private key in a file in JSON format. Ensure that you have the required permissions to access and use the file.

Upload File for Replication Manager to use the credentials in it to access data.

Add Cloud Credential   ✕

Cluster

Select...   ⌄

Cloud Storage Type

GCS   ⌄

Name *

Enter a unique name for the cloud credential

Input Type

File   ⌄

Service Account Private Key (json format) *

Upload File

Cancel   Save

6. Click Save.

Results

The GCP cloud credentials appear on the **Cloud Credentials** page. You can use these credentials when you create the replication policies. The credentials allow Replication Manager to access the cloud data in the source cluster.

Add IDBroker to use temporary AWS session credentials

You can use temporary AWS session credentials through IDBroker to provide just-in-time, minimum required access to replicate data using replication policies. You must complete a few prerequisites before you configure IDBroker to use in replication policies. You can then add the credentials in Cloudera Replication Manager. Alternatively, you can add an external account for the IDBroker topology in Cloudera Manager.

Before you use temporary AWS session credentials in a replication policy, you must:

1. have an AWS account with an IAM role that has the required permissions to access the target S3 bucket and has the necessary trust relationships set up.
2. install a role instance for IDBroker
3. configure non-HA IDBroker on the Cloudera Base on premises cluster.
4. add the cloud credential in Cloudera Replication Manager.

   Alternatively, you can add an external account for the IDBroker topology in Cloudera Manager.

How temporary AWS credentials for replication policies works

Some deployments require temporary AWS session credentials to provide just-in-time, minimum required access to replicate data using replication policies. You can achieve this task using IDBroker. You can use temporary AWS credentials, through the IDBroker service, to replicate HDFS data, Hive external tables, and HBase data from Kerberized Cloudera Base on premises 7.1.9 SP1 clusters or higher using Cloudera Manager 7.11.3 CHF7 or higher versions to S3 buckets using Cloudera Replication Manager.

You can also use the temporary AWS credentials to replicate the HDFS data from S3 buckets to Kerberized Cloudera Base on premises 7.1.9 SP1 clusters or higher using Cloudera Manager 7.11.3 CHF7 or higher versions.

IDBroker is a REST API built as an extension of Apache Knox's authentication services. It allows an authenticated and authorized user to exchange a set of credentials or a token for short-lived cloud vendor access tokens.

To acquire the temporary AWS credentials, you create an IDBroker topology and then map the Kerberos users (or groups) to an AWS IAM Role. During the replication policy run, Replication Manager invokes IDBroker, and the IDBroker then uses the mapping between the on-premises Kerberized user and the IAM Role to request an AWS session token for that role.

Use case

> An organization uses the same on-premises cluster across all their departments, and each department has its own AWS account so that it can replicate its required data from the on-premises cluster to its own AWS account when necessary. Each department depending on their requirements might either want to leverage the cloud storage capabilities to store data, or use the cloud processing capabilities to run workloads, analyze the data, or any other purposes.

Authentication methods to use AWS credentials in replication policies

You can choose long-term AWS cloud credentials or temporary AWS session credentials when you want to replicate HDFS data, Hive external tables, and HBase data from Cloudera Base on premises clusters to S3 buckets on Cloudera on cloud.

Long-term cloud credentials

You can use long-term credentials to replicate data to the cloud using replication policies. To use long-term cloud credentials in a replication policy, you must:

- have an AWS account, and access key and secret key for it.
- register an external account in Cloudera Manager using AWS access key and AWS secret key.

  You can add an external account on the Cloudera Manager Administration External Accounts page. The external account serves as an authentication method during data replication, using replication policies, from Cloudera Base on premises clusters to cloud.

- add the cloud credential in Cloudera Replication Manager.

The following use cases illustrate scenarios where you can use long-term AWS credentials:

- Environments where you have multiple users and multi-tenancy – In this instance, you can add an **Add Access Key Credentials** external account in Cloudera Manager for Cloudera Base on premises cluster, add the cloud credentials in the Cloudera Replication Manager, and then create a replication policy.
- Single user cluster, or where all the users of the cluster have the same privileges to the data in Amazon S3 – In this instance, you can add **IAM role-based authentication** in Cloudera Manager for Cloudera Base on premises cluster, add the cloud credentials in the Cloudera Replication Manager, and then create a replication policy.

Temporary AWS session credentials

You can use temporary AWS session credentials to provide just-in-time, minimum required access to replicate data using replication policies. Before you use temporary AWS session credentials in a replication policy, you must:

1. have an AWS account with an IAM role that has the required permissions to access the target S3 bucket and has the necessary trust relationships set up.
2. install and configure IDBroker on the Cloudera Base on premises cluster.
3. add the cloud credential in Cloudera Replication Manager.

   Alternatively, you can add an external account for the IDBroker topology in Cloudera Manager.

Adding a role instance to IDBroker in Cloudera Manager
To use IDBroker to access the cloud credentials, you must add a role instance to IDBroker, and then you configure the required properties for it in Cloudera Manager.

Before you begin
You must have an AWS user account that has:

- an IAM policy which allows you to access the S3 bucket to which you want to replicate the data.
- an IAM role, which you can assume, that has the above policy attached to it.

If you have upgraded or installed the Cloudera Base on premises cluster, ensure that IDBroker is available. IDBroker is automatically configured by Cloudera Manager in Cloudera deployments, where Knox is installed. For more information about IDBroker, see Configuring access to S3.

Procedure

1. Go to the source Cloudera Manager Clusters *KNOX SERVICE* Instances page.

   ⚠️ **Important:** If you are replicating HDFS data from cloud to Cloudera Base on premises cluster, add the role instance on the target Cloudera Manager.

2. Click Actions Add Role Instances .
3. Select the required **Knox IDBroker** host on the **Add Role Instance to Knox** modal window to install it to the required host, and then click Continue.
4. Specify an IDBroker Master Secret, and click Finish.
   The **Instances** tab shows the added role instance.

What to do next
Configure the required properties for IDBroker to use in replication policies.
Configuring IDBroker to use in replication policies
After you add a role instance to install IDBroker, you configure the required properties for it in Cloudera Manager for the Cloudera Base on premises cluster.

Procedure

1. Go to the source Cloudera Manager Clusters *KNOX SERVICE* Instances page.

   ⚠ **Important:** If you are replicating HDFS data from cloud to Cloudera Base on premises cluster, add the role instance on the target Cloudera Manager.

2. Click the role instance on the **Instances** tab to open the Knox IDBroker service.

3. Click the **Configuration** tab.

4. Click **Continue Editing Role Instance** if an alert appears.

5. Add the Kerberos username and ARN in the *[***USERNAME***]=[***ARN***]* format in the Knox IDBroker AWS User Mapping property. This maps the Kerberos user to the AWS role.

   If you have multiple usernames, you can map it as shown in the following sample Knox IDBroker AWS User Mapping value. The sample shows Kerberos users *repl*, *hdfs*, and *hbase* mapped to an AWS IAM role:

   repl=arn:aws:iam::134232123254:role/cldr-demo-role;hdfs=arn:aws:iam::134232123254:role/cldr-demo-role;hbase=arn:aws:iam::134232123254:role/cldr-demo-role

   📝 **Note:** You can map multiple Kerberos users and groups to the same AWS IAM role, but one Kerberos user cannot have multiple AWS IAM roles mapped to it.

6. Add the following in the Kerberos Proxy Block property for HDFS replication policies. This allows the hdfs user to impersonate the Kerberos user during the replication policy run:

   • "hadoop.proxyuser.hdfs.groups": "*[***KERBEROS USER***]*
   • "hadoop.proxyuser.hdfs.hosts": "*"

7. Perform the following steps for HBase replication policies:

   a) Add the following details in the Kerberos Proxy Block property. This allows the hbase users to impersonate the Kerberos users during the replication policy run.

      • "hadoop.proxyuser.hbase.groups": "*[***KERBEROS USER***]*
      • "hadoop.proxyuser.hbase.hosts": "*"

   b) Add the following details in the Cluster-wide Advanced Configuration Snippet (Safety Valve) for core-site.xml in HDFS configuration. This allows the hbase users to impersonate the Kerberos users when exporting the initial snapshot from HDFS.

      • "hadoop.proxyuser.hbase.groups": "*[***KERBEROS USER***]*
      • "hadoop.proxyuser.hbase.hosts": "*"

8. Configure the IDBroker Knox Token TTL property to ensure that the configured Knox session token time is greater than the time required to complete a replication policy run, and Save Changes.

9. Use the default aws-cab topology, or create a custom topology, if required, using the Knox IDBroker Advanced Configuration Snippet (Safety Valve) for conf/cdp-resources.xml property. You can also create multiple topologies depending on your use case requirements.

   The following sample code shows a custom topology added to the Knox IDBroker Advanced Configuration Snippet (Safety Valve) for conf/cdp-resources.xml property:

```
<property>
    <name>[***TOPOLOGY1***]</name>
    <value>
        providerConfigRef=cab-providers#IDBROKER:cloud.policy.config.prov
ider=default#IDBROKER:cloud.client.provider=AWS
    </value>
</property>
```

10. Complete the following steps to create the aws.credentials.key and aws.credentials.secret aliases in the topology.

   a) Search for the Save Alias Command Input property:
   b) Enter *[***TOPOLOGY1***]*.aws.credentials.secret=*[***SECRET***]*, and click Save Changes.
   c) Click Actions Save Alias - IDBroker .
   d) Enter *[***TOPOLOGY1***]*.aws.credentials.key=*[***ACCESS KEY***]*, and click Save Changes.
   e) Click Actions Save Alias - IDBroker .

11. Add the following credential details to use the default AWS topology in IDBroker if all the required IAM roles are assumed by a single set of long-term AWS keys. IDBroker uses these credentials to authenticate and to request session tokens from AWS Session Token Service (AWS STS). These credentials are used by IDBroker only to request session tokens and are not used during replication.

   •   Knox IDBroker AWS Credentials Key
   •   Knox IDBroker AWS Credentials Secret

12. Save the changes.

13. Restart Stale Services, if any.

What to do next
You can add the cloud credential in Cloudera Replication Manager. Alternatively, you can add an external account for the IDBroker topology in Cloudera Manager to use in replication policies.
Adding IDBroker credentials in Cloudera Replication Manager
After you install and configure non-HA IDBroker in Cloudera Manager, you add the cloud credentials in Cloudera Replication Manager. Alternatively, you can create an IDBroker-based external account in Cloudera Manager to use AWS temporary credentials for data replication using replication policies.

Before you begin
Ensure that you have completed the steps to install and configure IDBroker in Adding a role instance to IDBroker in Cloudera Manager on page 109.

Procedure

1. Go to the Cloudera Replication Manager Cloud Credentials page.

2. Click Add.

3. Enter the following details on the **Add Cloud Credential** modal window:

Option	Description
Cluster	Choose the source Cloudera Base on premises cluster where you have configured the IDBroker mapping. ⚠️ **Important:** If you are replicating HDFS data from cloud to Cloudera Base on premises cluster, the IDBroker must be configured on the target Cloudera Manager.
Cloud Storage Type	Choose S3.
Name	Provide a name for the IDBroker topology.
Authentication Type	Choose IDBroker Topology.
IDBroker Address	Enter the IDBroker host and port details in the https://*[***IDBROKER HOST***]*:*[***IDBROKER PORT***]*/ gateway format. To identify the required IDBroker address to add in this field, go to the source Cloudera Manager Clusters *KNOX SERVICE* Instances page. The Hostname for the Knox IDBroker Role Type is the IDBroker address. For example, `https://cldrcld-1.cld.root.site:8444/ gateway`.

Option	Description
Cloud Topology	Enter the topology name.
	Enter the topology name you added in the Knox IDBroker Advanced Configuration Snippet (Safety Valve) for conf/cdp-resources.xml property on the source Cloudera Manager Clusters Knox service Configuration tab.

Results

The credentials appear on the **Cloud Credentials** page.

What to do next

You can now use the temporary AWS session credentials through the IDBroker credentials when you create HDFS, Hive, or HBase replication policies in Cloudera Replication Manager.

Adding and managing an IDBroker-based external account in Cloudera Manager

You can create an IDBroker-based external account in Cloudera Manager to use AWS temporary credentials for data replication using replication policies. Ensure that you have configured the required properties for it in Cloudera Manager for the Cloudera Base on premises cluster. Alternatively, you can add the IDBroker credentials in Cloudera Replication Manager.

Procedure

1. Go to the source Cloudera Manager Administration Settings AWS Credentials page.

> ⚠ **Important:** If you are replicating HDFS data from cloud to Cloudera Base on premises cluster, add the role instance on the target Cloudera Manager.

2. Complete the following steps to add an IDBroker-based external account in Cloudera Manager:

   a) Click Add IDBroker Topology for Authentication.

   The **Add IDBroker Topology for Authentication** modal window appears.

   b) Provide a Name for the IDBroker topology.

   c) Enter one or more comma-separated list of IDBroker addresses. The IDBroker address includes the IDBroker host and port details in the https://[****IDBROKER HOST***]:[***IDBROKER PORT***]/gateway format.

   To identify the required IDBroker address to add in this field, go to the source Cloudera Manager Clusters *KNOX SERVICE* Instances page. The Hostname for the Knox IDBroker Role Type is the IDBroker address.

   For example, https://cldrcld-1.cld.root.site:8444/gateway.

   d) Enter the Cloud topology or IDBroker topology to use for cloud connections.

> 💡 **Tip:** Enter the topology name that you added in the source Cloudera Manager Clusters Knox service Configuration Knox IDBroker Advanced Configuration Snippet (Safety Valve) for conf/cdp-resources.xml   property.

3. Perform one or more of the following steps to manage an IDBroker-based external account:

   a) Go to the source Cloudera Manager Administration Settings AWS Credentials page.

> ⚠ **Important:** If you are replicating HDFS data from cloud to Cloudera Base on premises cluster, add the role instance on the target Cloudera Manager.

   b) Click Actions Edit Credential to edit the credentials for the required **IDBroker Topology**.

   c) Click Actions Remove the credentials to remove the credentials for the required **IDBroker Topology**.

   d) Click Actions Edit Connectivity to open the **Connect to Amazon Web Services** modal window to view more details.

Ports for Cloudera Replication Manager

Before you create replication policies, you must ensure that the required ports are open and available for data replication. You can verify the mandatory ports using the Replication Manager network security diagram.
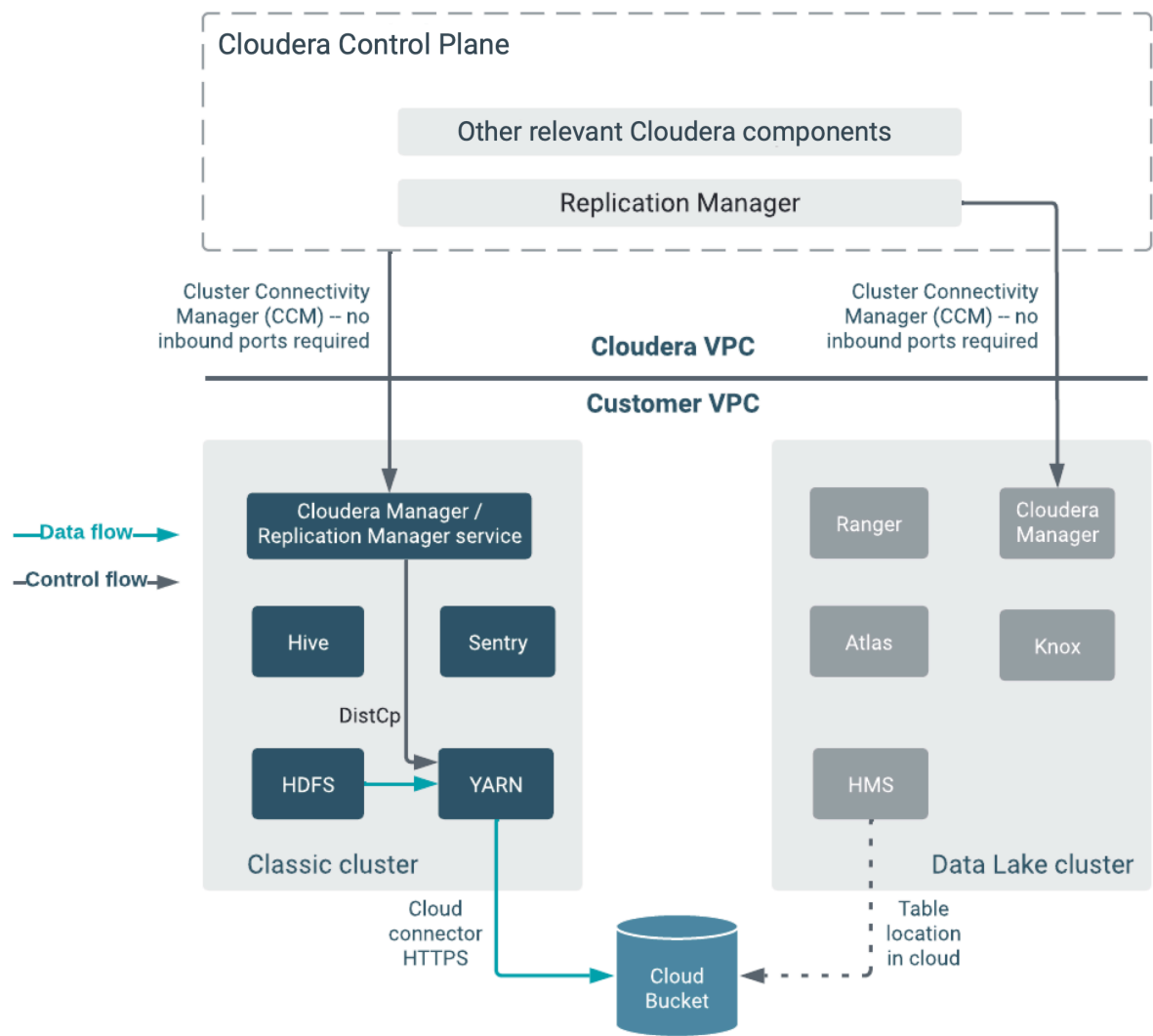
HDFS replication polices

The following ports must be open and available for Replication Manager for HDFS replication policies:

Table 10: Minimum ports required for HDFS replication policies

Connectivity required	Default Port	Type	Description
Data transfer from classic cluster hosts to cloud storage	80 or 443 (TLS)	Outbound	Outgoing port. All classic cluster nodes must be able to access S3/ ADLS Gen2 endpoint. **Note:** Before you create replication policies, ensure that the best practices for Cloudera on cloud on Microsoft Azure ADLS Gen2 (ABFS) are complete.
Classic cluster	6000-6049 for CCMv1 443 for CCMv2	Outbound	Connecting source classic cluster to the Cloudera Management Console through Cluster Connectivity Manager. For more information, see Outbound network access for Cluster Connectivity Manager, and Cluster Connectivity Manager overview.

The following system architecture diagram shows the interaction between components during HDFS replication using HDFS replication policies:

Figure 7: System architecture diagram for HDFS replication in Cloudera Replication Manager

Hive replication polices

The following ports must be open and available for Replication Manager for Hive replication policies:
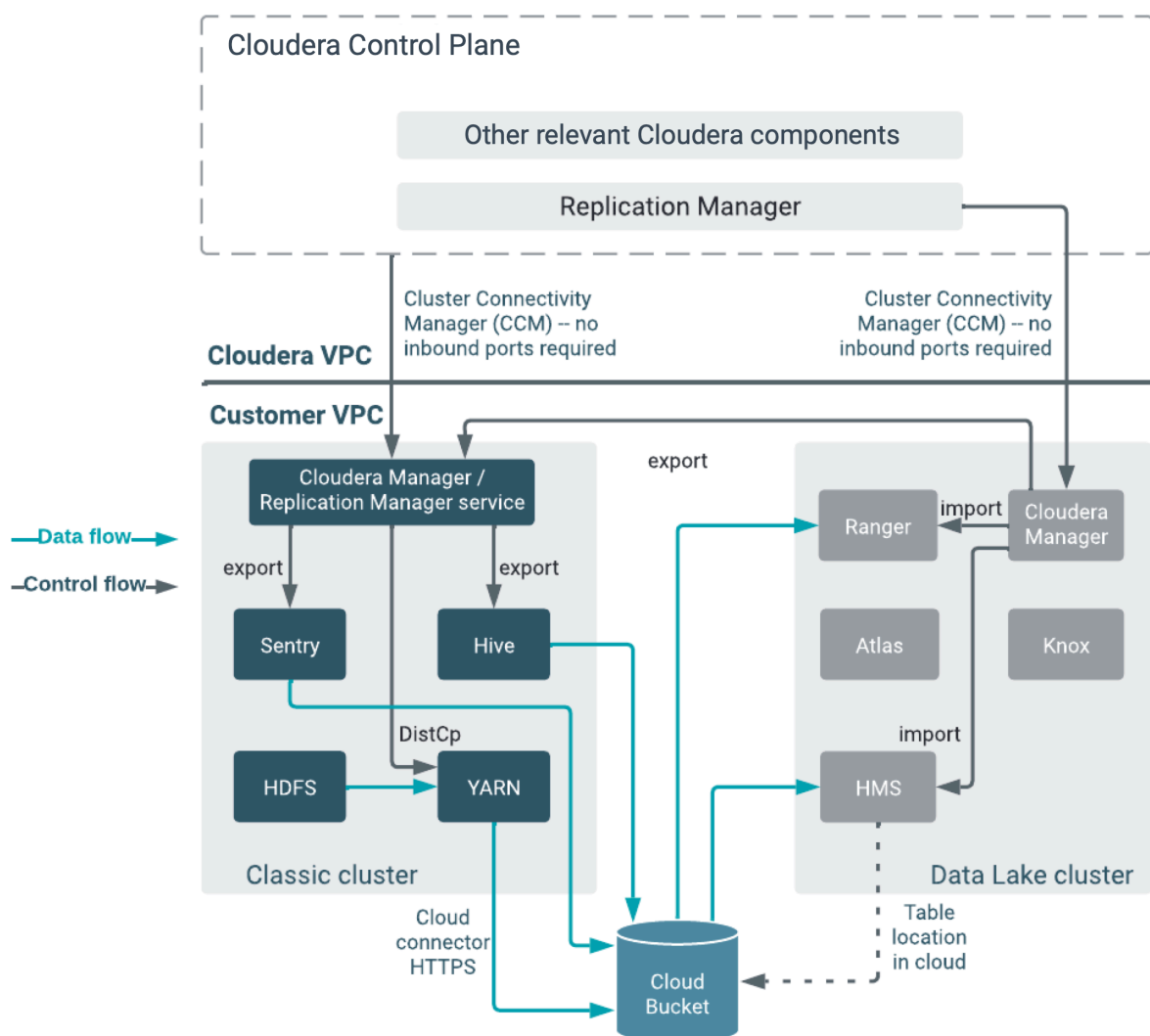
Table 11: Minimum ports required for Hive replication policies

Connectivity required for	Default Port	Port opened toward	Description
Data transfer from classic cluster hosts to cloud storage	80 or 443 (TLS)	Outbound towards cloud provider public hosts (For example, AWS)	Outgoing port. All classic cluster nodes must be able to access S3/ADLS Gen2 endpoint. **Note:** Before you create replication policies, ensure that the best practices for Cloudera on cloud on Microsoft Azure ADLS Gen2 (ABFS) are complete.

Connectivity required for	Default Port	Port opened toward	Description
Cloudera Manager Admin Console HTTP	7180 or 7183 (when TLS enabled)	Inbound towards destination Cloudera on cloud Cloudera Manager host	Incoming port. Open on the source cluster to enable the target Cloudera Manager in cloud to communicate to the on-premises Cloudera Manager.
Classic cluster	6000-6049 for CCMv1 443 for CCMv2	Outbound towards Cloudera Control Plane (Cluster Connectivity Manager and cluster proxy)	Connecting the source classic cluster to the Cloudera Management Console through Cluster Connectivity Manager For more information, see Outbound network access for Cluster Connectivity Manager, and Cluster Connectivity Manager overview.

The following system architecture diagram shows the interaction between components during Hive replication using Hive replication policies:

Figure 8: System architecture diagram for Hive replication in Cloudera Replication Manager

HBase replication policies

The following ports must be open and available for Replication Manager for HBase replication policies:

Table 12: Minimum ports required for HBase replication policies

Service	Ports	Ports opened toward	Description
Destination hosts of the AWS cluster or ADLS cluster (target cluster), and the Cloudera Manager server port on the source cluster	2181 and 16020	Outbound in destination Zookeeper and HBase RegionServer host(s)	Verify whether the ports 16020 for worker security group and 2181 for worker, master, and leader groups are open for connection from the source cluster to the destination cluster on AWS or Azure. This ensures that the source HBase service can communicate with Zookeeper and HBase services on the destination hosts uninterruptedly. For more information, see Ports for HBase replication.
HMaster	16000	Outbound in destination HBase Master host(s)	Open the port on the Master Nodes (HBase Master Node and any back-up HBase Master node). Before you select the Validate Replication option during the first HBase replication policy creation between two specific clusters, you must ensure that the port is open on the target cluster.
Cloudera Manager Admin Console HTTP	7180 or 7183	Inbound towards destination Cloudera on cloud Cloudera Manager host	Open on the source cluster to enable Data lake Cloudera Manager to communicate to the on-premises Cloudera Manager. Connects to destination SDX Data Lake Cloudera Manager.
Cluster Connectivity Manager for Cluster Connectivity Manager v1	6000-6049	For Cloudera Control Plane (Cluster Connectivity Manager and Cluster Proxy)	Required for SSL connections to the Cloudera Control Plane via Cluster Connectivity Manager to communicate with Replication Manager.
Data transfer from secondary node for AWS / ADLS Gen2 for Cluster Connectivity Manager v2	80 or 443	Outbound towards data transfer from secondary node for AWS / ADLS Gen2	Outgoing port. Open on all the HDFS nodes for AWS and ADLS Gen2. **Note:** Before you create replication policies, ensure that the best practices for Cloudera on cloud on Microsoft Azure ADLS Gen2 (ABFS) are complete.
Data Lake cluster	8443	On destination Cloudera on cloud Cloudera Manager/Knox host. (applicable when Knox is available on the on-premises source cluster)	Outgoing port. Configure the port on the Data Lake cluster as the outgoing port for Cloudera Management Console to communicate with Cloudera Manager and Knox.

Figure 9: System architecture diagram for HBase replication in Cloudera Replication Manager
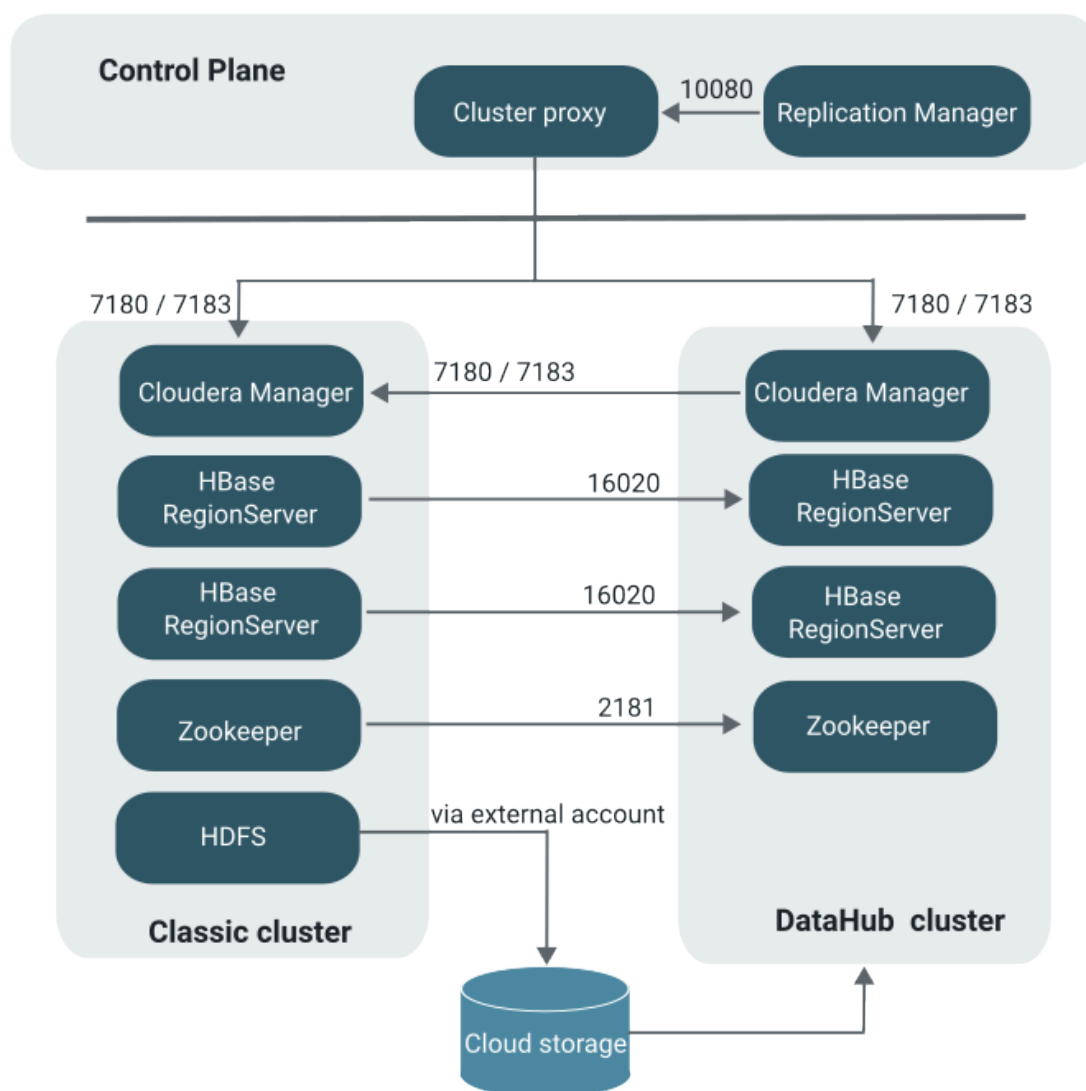
Best practices

Consider the following best practices while using Cloudera on cloud on Microsoft Azure ADLS Gen2 (ABFS):

- Ensure that the on-premises cluster (port 443) can access the *https://login.microsoftonline.com* endpoint. This is because the Hadoop client in the on-premises cluster (CDH/Cloudera Base on premises) connects to the endpoint to acquire the access tokens before it connects to Azure ADLS storage. For more information, see the *General Azure guidelines* row in the Azure-specific endpoints table.
- Ensure that the steps mentioned in the *General Azure guidelines* and *Azure Data Lake Storage Gen 2* rows in the Azure-specific endpoints table are complete so that the endpoint connects to the target path successfully.

Ports required for HBase replication policies

Open the ports 2181 and 16020 on the source and destination secondary nodes to ensure that the source HBase service can reach Zookeeper and HBase services on the destination hosts.

Use one of the following methods to open the required ports for HBase replication:

- Choose a security group for your environment and open the ports manually. In this method, you choose the security groups that are automatically created for the environment. By default, the security groups do not have any rules for Zookeeper and HBase ports, therefore, you must open the required ports manually after you create a Data Hub.

  After you open the ports, the required security groups are assigned to the nodes when the nodes are autoscaled. This is a one-time process that you must perform when you create a Cloudera Data Hub cluster.
- Define a security group with the required ports open, and assign it to the new Cloudera Data Hub environment. In this method, you define a security group for a VPC that contains inbound rules to open the required ports which include Zookeeper and HBase ports. When you create an environment, you assign this security group to it. If required, you can assign different security groups to the gateway node and other nodes.

  This method allows you to reuse the security groups in other new Cloudera Data Hub clusters. Security issues do not appear because the nodes in the same security group do not access each other by default. However, if required, you can add a separate rule to impose this restriction. Sharing the same security group for inbound and outbound network access rules remains as strict as having separate security groups for each environment, but the extra rules for Zookeeper and HBase ports do not need to be added at each environment creation.

The following use cases illustrate the situations where a requirement for autoscaling nodes during HBase replication might appear:

- You replicate HBase data to another Cloudera account or region in the same cloud provider. In this use case, ensure that the VPC/VNET peering is complete before you open the ports to establish connection over private networks.
- You replicate HBase data to Cloudera Operational Database or Cloudera Data Hub using a direct connection. In this use case, you ensure that public IPs and Zookeeper ports are not open to the internet.