Cloudera Replication Manager

CDP Public Cloud Replication Manager use cases

Date published: 2022-08-11
Date modified: 2024-02-26

https://docs.cloudera.com/

Legal Notice

Contents

Use cases for Replication Manager in CDP Public Cloud

Replication Manager is a service in CDP Public 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 CDP Private Cloud Base (version 7.1.4 and higher) clusters (HDFS, Hive external tables, and HBase data) to CDP Public Cloud clusters. You can also replicate HDFS data from classic clusters (CDH, CDP Private Cloud Base, and HDP clusters) to Public Cloud buckets. The supported Public Cloud services include Amazon S3 and Microsoft Azure ADLS Gen2 (ABFS).

You can use Replication Manager for a variety of use cases. Some major use cases include:

- Implementing a complete backup and restore solution.

  You might want to implement a backup and restore solution for HDFS data or Hive external tables. You create the replication policy based on the type of data you want to backup and restore. To implement this use case, you back up data in ClusterA to ClusterB. When the need arises, you can create another replication policy to restore the data from ClusterB to ClusterA.
- Migrating legacy data (for example, from CDH clusters) to CDP Public Cloud clusters on Amazon S3 and Microsoft Azure ADLS Gen2 (ABFS).

  Replication Manager supports replicating HDFS data, Hive external tables, and HBase data from CDH (version 5.13 and higher) clusters (HDFS, Hive, and HBase data) to CDP Public Cloud clusters. Before you create replication policies, you must prepare the clusters and complete the prerequisites.

  Alternatively, you can use CDP CLI to create replication policies to replicate HDFS data and Hive external tables.
- Replicating the required data to another cluster to run workloads or analytics.

  Sometimes, you might want to move workloads, especially heavy-duty workloads to another cluster to reduce the load and optimize the performance of the primary cluster, or run analytics on the required data on another cluster because you do not want to overload the primary cluster. In such scenarios, you can create replication policies in Replication Manager to replicate the required data at regular intervals. After replication, you can use the required tools to analyze the data.

Replicating multiple large tables using CDP Public Cloud Replication Manager

You can create multiple replication policies to replicate multiple large HDFS directories, Hive external tables, or HBase tables using Replication Manager in CDP Public Cloud.

When you have multiple tables or directories to replicate, you might want to create more than one replication policy to replicate the data. This is because even if one replication policy fails, it would not stop the replication process. You can troubleshoot and restart the failed replication policy or create a new replication policy for the tables or directories in the failed replication policy.

Scenario and solution

You want to replicate multiple large HBase tables (around 6000 tables) using the incremental approach. In this approach, you replicate data in batches. For example, 500 tables at a time. This approach ensures that the source cluster is healthy because you replicate data in small batches.

The following steps explain the incremental approach in detail:

1. You create an HBase replication policy for the first 500 tables.

   Internally, Replication Manager performs the following steps:

   a. Creates a disabled HBase peer on the source cluster or disables if the HBase peer existed.
   b. Creates a snapshot and copies it to the target cluster.

      HBase replication policies use snapshots to replicate existing HBase data; this step ensures that all the existing data existing is replicated.
   c. Restores the snapshot to appear as the table on the target.

      This step ensures that the existing data is replicated to the target cluster.
   d. Deletes the snapshot.

      The Replication Manager performs this step after the replication is successfully complete.
   e. Enables table's replication scope for subsequent replication.
   f. Enables the peer.

      This step ensures that the accumulated data is completely replicated.

   After all the accumulated data is migrated, the HBase service continues to replicate new/changed data in this batch of tables automatically.
2. Create another HBase replication policy to replicate the next batch of 500 tables after all the existing data and accumulated data of the first batch of tables is migrated successfully.
3. You can continue this process until all the tables are replicated successfully.

In an ideal scenario, the time taken to replicate 500 tables of 6 TB size might take around four to five hours, and the time taken to replicate the accumulated data might be another 30 minutes to one and a half hours, depending on the speed at which the data is being generated on the source cluster. Therefore, this approach uses 12 batches and around four to five days to replicate all the 6000+ tables to Cloudera Operational Database (COD).

For more information about the cluster versions supported by Replication Manager, clusters preparation process to complete before you create the HBase replication policy, HBase replication policy creation, and for an in depth analysis of the use case, see related information.

Related Information

Using HBase replication policy

Preparing to create an HBase replication policy

Creating an HBase replication policy

Tracking, monitoring, and troubleshooting replication policies

After you create a replication policy, you can track the replication policy job progress, monitor the status of the replication policies, and troubleshoot the failed replication policies in CDP Public Cloud Replication Manager.

Tracking and monitoring replication policies

You can view, track, and monitor the available replication policies on the Replication Manager UI:

- **Overview** page. This page shows the statistics in the Policies, Jobs, and Issues & Updates panels.

  Tip: The **Clusters** panel and page shows information about the available classic clusters, Data Lakes, and Data Hubs to use in Replication Manager

- **Replication Policies** page. This page lists all the replication policies and information about each replication policy.

Troubleshooting replication policies

You can use one of the following methods to troubleshoot a failed replication policy:

- Click the failed job on the Replication Manager Replication Policies Job History panel. The errors for the failed job appear.
- Click Cloudera Manager Running Commands for the source or target cluster. The recent command history shows the failed commands.
- Open the service logs in Cloudera Manager to track the errors on the source cluster and target cluster. For example, HBase service logs.

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

Related Information

Overview page
Replication Policies page
Using HDFS replication policies
Using Hive replication policy
Using HBase replication policy
Troubleshooting replication policies in CDP Public Cloud