

Sidecar migration from CDH to CDP

Date published: 2022-02-14

Date modified:



Legal Notice

© Cloudera Inc. 2024. All rights reserved.

The documentation is and contains Cloudera proprietary information protected by copyright and other intellectual property rights. No license under copyright or any other intellectual property right is granted herein.

Unless otherwise noted, scripts and sample code are licensed under the Apache License, Version 2.0.

Copyright information for Cloudera software may be found within the documentation accompanying each component in a particular release.

Cloudera software includes software from various open source or other third party projects, and may be released under the Apache Software License 2.0 (“ASLv2”), the Affero General Public License version 3 (AGPLv3), or other license terms. Other software included may be released under the terms of alternative open source licenses. Please review the license and notice files accompanying the software for additional licensing information.

Please visit the Cloudera software product page for more information on Cloudera software. For more information on Cloudera support services, please visit either the Support or Sales page. Feel free to contact us directly to discuss your specific needs.

Cloudera reserves the right to change any products at any time, and without notice. Cloudera assumes no responsibility nor liability arising from the use of products, except as expressly agreed to in writing by Cloudera.

Cloudera, Cloudera Altus, HUE, Impala, Cloudera Impala, and other Cloudera marks are registered or unregistered trademarks in the United States and other countries. All other trademarks are the property of their respective owners.

Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES NOT WARRANT THAT CLOUDERA PRODUCTS NOR SOFTWARE WILL OPERATE UNINTERRUPTED NOR THAT IT WILL BE FREE FROM DEFECTS NOR ERRORS, THAT IT WILL PROTECT YOUR DATA FROM LOSS, CORRUPTION NOR UNAVAILABILITY, NOR THAT IT WILL MEET ALL OF CUSTOMER’S BUSINESS REQUIREMENTS. WITHOUT LIMITING THE FOREGOING, AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, CLOUDERA EXPRESSLY DISCLAIMS ANY AND ALL IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, QUALITY, NON-INFRINGEMENT, TITLE, AND FITNESS FOR A PARTICULAR PURPOSE AND ANY REPRESENTATION, WARRANTY, OR COVENANT BASED ON COURSE OF DEALING OR USAGE IN TRADE.

Contents

Sidecar migration from CDH to CDP.....	4
Process Overview.....	4
Pre-migration steps.....	5
Planning.....	5
Preparing your hosts.....	5
Replicate configurations.....	5
Set up TLS.....	6
Set up Kerberos.....	6
Migrating Services.....	6
Sentry to Ranger Sidecar migration to CDP.....	6
Navigator to Atlas Sidecar migration from CDH to CDP.....	7
HDFS Sidecar Migration to CDP.....	8
Hive Sidecar Migration from CDH to CDP.....	8
YARN Fair Scheduler to Capacity Scheduler Sidecar migration from CDH to CDP.....	8
HBase Sidecar migration from CDH to CDP.....	9
Impala Sidecar migration to CDP.....	9
Kudu Sidecar Migration from CDH to CDP.....	9
Solr Sidecar migration from CDH to CDP.....	10
Migrating from CDH 5.....	10
Migrating from CDH 6.....	10
Oozie Sidecar migration to CDP.....	10
Accumulo Sidecar migration to OpDB.....	11
Post-migration steps.....	11
Sidecar migration for Cloudera Flow Management.....	11
Testing applications.....	12

Sidecar migration from CDH to CDP

“Side-car” migration allows you to migrate an existing CDH deployment to a new cluster deployed on new (or different) hardware running the latest version of CDP Private Cloud Base. Migrations completed using this process have several advantages, including minimizing downtime during the migration process. You also will have the ability to test your workloads at production scale on the new cluster while keeping your production system running until you determine that the new cluster is ready.

Cloudera provides a variety of tools and documentation to help you move your data, service and cluster configurations, security configurations, and security policies to the new cluster. The exact procedures to follow may depend on the mix of services deployed on your cluster. This document can help you to learn how to plan and sequence your upgrade. You may need to engage Cloudera Professional Services to plan and execute some parts of the upgrade that could require special procedures not covered in this document.

This document will also provide you with links to useful information about what you need to know about changes in Cloudera Runtime components and how those might affect your workloads. Cloudera recommends that you review those documents closely to avoid unexpected changes in behavior that may delay your migration.

This method of upgrading your cluster has several advantages:

- Service-level agreements for individual workloads are easier to meet because the legacy versions and the CDP versions can run in parallel without impacting each other.
- Single, well-contained tenants can move one at a time without requiring a single coordinated event across all tenants.
- Rollback requires coordination only at the workload or tenant level rather than at the whole cluster level.
- Enables moving directly to CDP from any CDH 5 or CDH 6 version.
- Sidecar migration provides the ability to test your workloads at production scale on the new cluster without impacting your live production environment.

You can also use a “rolling side-car migration”, where you decommission some of the existing cluster hosts, install CDP on the decommissioned hosts, and then migrate workloads one at a time. As workloads are migrated, you again decommission more hosts and repeat the process.

For an overview of other upgrade and migration options, see [The Four Upgrade and Migration Paths to CDP from Legacy Distributions](#).

Process Overview

A side-car migration consists of the following phases, which are discussed in detail in other topics.

1. Pre-migration and planning

In this phase, you analyze your workloads and datasets to determine the type and quantity of hosts required for the migration. You also need to consider changes and upgrades to various components used in the source cluster. Some components are no longer available in CDP and you will need to understand how to migrate to services that provide similar functionality.

After determining the specifications for the destination cluster, you can begin to provision your hosts with a supported operating system and install the CDP Private Cloud Base software, which includes Cloudera Manager and Cloudera Runtime.

2. Migration

In this phase, you migrate your data and workloads to the destination cluster and begin testing for functionality and performance. After migrating your data, you can configure the replication process to use “snapshots” so that you

can keep the data on the destination cluster synchronized with the source cluster as you complete your workload migration and testing.

If the cluster is using HDFS Transparent Data Encryption, you may also need to migrate KMS ACLs to the new CDP Cluster.

If the source cluster includes Apache Sentry you will need to migrate the Sentry service to Apache Ranger. If the source cluster includes Cloudera Navigator you will also need to migrate to Apache Atlas.

3. Post-migration

Some components require additional steps that must be performed after the data and workload migration are completed. Review the sections for your components in this publication to learn more about these requirements. When these steps are complete and you have completed your testing, you can move your new cluster(s) into production and de-commission the workloads running on the source cluster.

Pre-migration steps

There are steps to take before you begin a Sidecar migration.

Planning

To plan your migration, you need to consider a variety of items.

- When you provision the new CDP deployment, you should consider the type and capacity of the hosts you deploy in your clusters. Ensure that hardware is sufficient in speed and capacity to handle your current and planned workloads.
- Depending on the software running on the source cluster, not all services deployed in CDH clusters have exact matches in CDP. Consider how your services will map to updated versions of those services or to new services designed to replace the legacy software. Configurations and performance tuning may need to be adjusted.
- Plan for any updates that may be required in your application code. Newer versions may have updated or deprecated APIs and replacement services could require you to modify or re-architect your applications to use the correct APIs.

Preparing your hosts

Procedure

1. Install a supported operating system on all hosts. To learn about supported operating systems, see the [Cloudera Support Matrix](#).
2. Install Cloudera Manager and Cloudera Runtime on the hosts. See the [CDP Private Cloud Base Installation Guide](#)

Replicate configurations

Before you begin

When you create your new CDP Private Cloud Base cluster(s) you may need to copy some of your custom/tuning configuration from the source cluster to the new cluster. These can include properties like thresholds, memory limits, ports numbers, directory locations, log file locations, etc. These configurations exist for services, clusters, and hosts. Note that many CDP services have been updated from earlier versions and there may not be a one-to-one correspondence of configuration parameters from the old service to the new service. You will need to ensure that you

understand the new set of configuration options in CDP and that your configurations are valid and appropriate for your workloads and environment in which the new cluster is deployed..

For migrations from CDH, Cloudera Manager can help you identify configuration properties that have been modified from their default settings so you can better assess what, if any, configuration settings should be explicitly brought over to the new CDP cluster:

Procedure

1. Open the Cloudera Manager Admin Console on the source deployment.
2. Click ConfigurationNon-default values.
A list of host and service properties and their modified values displays.
3. Open the Cloudera Manager Admin Console on the destination deployment.
4. Manually set the desired configuration values on services and/or hosts in the new cluster. For more information, see [Modifying Configuration Properties Using Cloudera Manager](#).
5. For services that have been updated or replaced, you will need to evaluate any functional or performance tuning that may be required in the new deployment. The CDP documentation has a great deal of information about setting up Cloudera Runtime services. Cloudera Professional Services can also assist you with tuning the new services.

If you are migrating the Hive service, see [Hive Sidecar Migration from CDH to CDP](#) on page 8 for additional guidance on configuring Hive in CDP.

Set up TLS

See:

- [Configuring TLS Encryption for Cloudera Manager Using Auto-TLS](#)
- [Manually Configuring TLS Encryption for Cloudera Manager](#)

Set up Kerberos

See [Configuring Authentication in Cloudera Manager](#)

Important: If you enable Apache Ranger on the cluster, you must configure Kerberos to enforce strong authentication of users. Running Apache Ranger on a non-Kerberized CDP environment is not supported.

Migrating Services

The steps required in this phase of a Sidecar migration can vary depending on the components deployed in the source cluster and planned for the destination cluster. In some cases, the order in which you complete these migrations should be considered due to dependencies among some components.

Sentry to Ranger Sidecar migration to CDP

You can migrate Apache Sentry to Apache Ranger by exporting permissions from the source cluster and then importing them to the destination cluster.

See the following topics:

- [Migrating from Sentry to Ranger](#)
- [Exporting Permissions from Sentry Server](#)
- [Ingesting permissions into Ranger](#)

Navigator to Atlas Sidecar migration from CDH to CDP

To migrate from Navigator to Apache Atlas, you can extract metadata from the source cluster and then transform and import the metadata on the destination cluster.

You can migrate your business metadata such as tags, custom properties (definitions and entity assignments), Managed metadata properties (definitions and entity assignments), original and updated entity names and descriptions.

You can also migrate technical metadata from Hive, Impala, Spark, and Referenced HDFS or S3.

Navigator audit data is not migrated to Atlas. You can, however, archive the Navigator audits on the original cluster, and Atlas will begin collecting audit data on the new cluster. See [Transitioning Navigator audits](#).

Entity metadata from unreferenced S3 and HDFS entities, and metadata for Sqoop, Pig, MapReduce, Oozie, and YARN, Navigator policies, Navigator configuration settings are not transitioned to Atlas.

To migrate the technical and business Navigator metadata, you run a command-line tool on the destination= cluster, pointing it to the source cluster that extracts the metadata from Navigator and then run another tool to convert the extracted data for use in Atlas.

1. If Navigator is configured with TLS/SSL, run the following commands on the destination cluster:

```
export KEYSTORE_PATH=<keystore-path-on-source-cluster>;
export KEYSTORE_PASSWORD=<keystore-password-on-source-cluster>;
export TRUSTSTORE_PATH=<truststore-path-on-source-cluster>;
export TRUSTSTORE_PASSWORD=<truststore-password-on-source-cluster>
```

2. Extract Cloudera Navigator data:

- a. Log in to the Cloudera Manager server host on the destination cluster.
- b. Ensure that the JAVA_HOME environment variable points to the JDK used by the host.
- c. Open the following file in a text editor:

```
/opt/cloudera/cm-agent/service/navigator/cnav.sh
```

- d. Run the the cnav.sh script using the following options:

```
/opt/cloudera/cm-agent/service/navigator/cnav.sh
-n <URL of Navigator server on source cluster>:7187
-u <Navigator username>
-p <Navigator Password>
-c <cluster name>
-o <path to output file>
```

3. Transform the metadata:

On the Cloudera Manager server, run the following command:

```
/opt/cloudera/parcels/CDH/lib/atlas/tools/nav2atlas/nav2atlas.sh
-f <output file name used in cnav.sh>
-o '/tmp/nav/nav2atlas/[***clustername***]_atlas_data.zip' clusterName
'['***clustername***']'
```

4. Start the Atlas service.

Atlas will automatically begin to import the entities.

For more information, see [Transitioning Navigator content to Atlas](#).

HDFS Sidecar Migration to CDP

You can migrate data stored in HDFS to CDP.

To migrate data stored in HDFS from CDH to CDP, use Replication Manager, available in Cloudera Manager. See ["HDFS Replication "](#)

Hive Sidecar Migration from CDH to CDP

If your cluster uses Apache Hive, you will need to migrate configurations, data, and workloads to the destination cluster.

Before you begin

CDP includes a new version of Hive, Hive 3. There are changes you should be aware of as you plan to migrate your Hive data to CDP.

Procedure

1. Implement configuration changes on the destination cluster.
 - a) Configure properties to the After Upgrade value documented in [Hive Configuration Property Changes](#) .
 - b) Configure critical Hive configurations documented. See [Customizing critical Hive configurations](#).
 - c) Set configuration overrides. See [Setting Hive Configuration Overrides](#).
 - d) Set required Hive Configuration Requirements and Recommendations documented in [Hive Configuration Requirements and Recommendations](#) .
 - e) If you configured high availability for the Hive Metastore in the source cluster, ensure that it is also enabled in the destination cluster. See [Configuring HMS for high availability](#).
2. Use Replication Manager to migrate your Hive tables and metastore from the source cluster to the destination cluster. Replication Manager is available from the Cloudera Manager Admin console on the destination cluster. See [Hive/Impala replication policy](#). You can begin replicating data before you are ready to take the new cluster to production and then use [Snapshots](#) to keep the destination cluster in sync with the data on the source cluster. If you check the box to replicate Sentry policy to Ranger, you cannot change the Ranger policy until your replications and migration are complete. See [Sentry to Ranger replication for Hive replication policies](#).
3. Migrate Workloads. See [Migrating Hive workloads from CDH](#).
4. Complete post-migration steps.
 - a) Read all the Hive documentation for CDP, which includes: Configuring HMS for high availability [Configuring HMS for high availability](#)
 - b) If you have any Impala workloads, see [Impala Sidecar Migration](#).
 - c) Read the Apache Ranger documentation to set up access to Hive tables: [Using Ranger to Provide Authorization in CDP](#)
 - d) Handle syntax changes between Hive 1/2 to 3 as documented in [Handling syntax changes](#)
 - e) Understand ACID tables, as documented in [Migrating Hive workloads to ACID](#)
 - f) Generate statistics as documented in [Generating statistics](#)
 - g) Convert HIVE CLI scripts to Beeline as documented in [Converting Hive CLI scripts to Beeline](#)

YARN Fair Scheduler to Capacity Scheduler Sidecar migration from CDH to CDP

The YARN Fair Scheduler is not supported in CDP. You must migrate to use the Capacity Scheduler.

Cloudera provides a tool to convert the Fair Scheduler into a Capacity Scheduler configuration. However, the features of Capacity Scheduler are not the same as the features of Fair Scheduler. Hence, the conversion utility cannot convert

every Fair Scheduler configuration into a Capacity Scheduler configuration. After the automatic conversion and once the migration is completed, you must manually tune the scheduler configurations to ensure that the resulting scheduling configuration fits your organization's internal goals and SLAs.

For a complete description of the process, and to learn more about the differences between Fair Scheduler and Capacity Scheduler, see [Fair Scheduler to Capacity Scheduler migration overview](#)

For the steps required to run the conversion utility (called fs2cs), see [Converting Fair Scheduler to Capacity Scheduler using the fs2cs conversion utility](#)

HBase Sidecar migration from CDH to CDP

You can use HBase replication and snapshots to migrate HBase tables to the destination cluster.

1. Prepare for the migration. Learn about changes in HBase and what actions you need to take before migrating your data. See [Preparing for data migration](#).
2. Use Cloudera Manager on the source cluster to take a snapshot of all tables you want to migrate to the destination cluster. See [Migrating HBase from CDH or HDP](#).
3. Export the snapshots to the destination cluster. See [Migrating HBase from CDH or HDP](#). For additional information and debugging advice, see:
 - [Exporting a Snapshot to Another Cluster \(CDH 5\)](#)
 - [Exporting a Snapshot to Another Cluster \(CDH 6\)](#)
4. Import the snapshot on the destination cluster by running the following command in the HBase shell:

```
restore_snapshot <snapshot name>
```

5. If you are migrating from CDH 5, you may need to update your client application code. See [Known Incompatibilities when Upgrading from CDH and HDP](#).
6. Review the following topics for additional incompatibilities with CDH 5:
 - [Remove PREFIX_TREE Data Block Encoding](#)
 - [Validate HFiles](#)
 - [Check co-processor classes](#)
7. Verify and validate the migrated data on the destination cluster. See [Verifying and validating if your data is migrated](#)
8. Test your applications or queries.
9. When you have verified and tested your application code and are ready to move the applications to production, repeat the above steps to synchronize the HBase data with the source cluster.

You can use the Hash Table and Sync Table tool to synchronize table data. This tool can help improve performance when synchronizing large tables. See [Use HashTable and SyncTable Tool](#).

Impala Sidecar migration to CDP

Side-car migration mechanism is best employed when you may have tighter service-level agreements that preclude an extended, multi-hour downtime for your workloads. This process aims to minimize downtime on individual workloads while providing a straightforward roll-back mechanism on a per-workload basis.

See [Impala Side-car migration](#)

Kudu Sidecar Migration from CDH to CDP

If your cluster uses Kudu, you will need to migrate Kudu data and configurations to the destination cluster.

Procedure

1. Back up all your data in Kudu using the kudu-backup-tools.jar Kudu backup tool.
2. Manually apply any custom Kudu configuration in your new cluster that you had in your old cluster.
3. Copy your backed up data to the target CDP cluster.
4. Restore your backed-up Kudu data using the Kudu backup tool.

What to do next

For details, see: [Migrating Kudu Data to CDP Private Cloud](#)

Solr Sidecar migration from CDH to CDP

There are several ways to migrate your Cloudera Search indexes to a new cluster.

Migrating from CDH 5

Procedure

Copy and re-index the data:

- a) Use Replication Manager in Cloudera Manager to copy all of the data you are indexing from the source to the destination cluster. See [Use Replication Manager to migrate to CDP Private Cloud Base](#).
- b) Re-index your collections. See [Reindexing Solr collections](#).

Migrating from CDH 6

About this task

The indexes used by Solr 7 in CDH 6 are compatible with Solr 8 that is included with CDP. In most cases, you can simply copy the data and indexes from the source cluster to the destination cluster. If you use the MRIT indexer to index your data, this will fail and you will need to recreate the indexes on the destination cluster. See [Reindexing Solr collections](#).

Oozie Sidecar migration to CDP

To migrate Oozie to CDP, you must manually migrate workloads and configurations to the destination cluster.

The Oozie database on the source cluster is automatically configured to purge data that is older than 30 days. If you need to retain this data, Cloudera recommends that you back up the data and make it available for running simple SQL queries. See [Back Up the Databases](#).

Ensure that you have set up a new Oozie database on the destination cluster. (You may have already done this when you installed Cloudera Manager on the destination cluster.)

Most Oozie workloads and configurations will need to be migrated manually to the destination cluster. Although you can copy the XML and property files from the source cluster, these files will require extensive modifications before they can be used in the destination cluster. Cloudera Professional Services can assist you with these migrations.

In general, you will need to ensure that the following settings are correct for the destination cluster:

- Name-node settings for each workflow.
- Resource-manager settings for each workflow.
- Paths, if the locations are being changed
- JDBC URLs, including Hive must point to the new database,
- DFS paths may need to be replaced with S3, Azure, as appropriate
- New cluster-specific configurations for the Workflow (for example, Amazon S3 credentials.)

- Coordinator Jobs that have future sdf runs scheduled. Recreate these jobs to match what was included in CDH. For such Jobs, the Workflow and its configuration might be stored in Hue.
- Copy any custom sharelib jars from the source to the destination cluster. The jars are stored here: `/user/oozie/share/lib/lib_{TIMESTAMP}/{COMPONENT}`
- If you have custom application jars used by any of your workflows, please make sure you recompile them with the 3rd-party libraries used by the latest Cloudera runtime. For example on older versions of CDH, Oozie was using Guava 11, but in CDP 7 it is using Guava 28 or higher. If you compiled your application with Guava 11 you must also upgrade Guava.
- On the destination cluster, update the sharelib using the following command:

```
oozie admin -oozie {URL} -sharelibupdate
```

Accumulo Sidecar migration to OpDB

You can migrate the Accumulo service to Operational Database powered by Apache Accumulo (OpDB) on CDP

Before you begin

Follow these steps to migrate the Accumulo service's configuration and data to Operational Database powered by Apache Accumulo (OpDB).

Procedure

1. Download and install the Operational Database powered by Apache Accumulo CSD and install it on the destination cluster. See [Install Operational Database powered by Apache Accumulo CSD file](#).
2. Ensure that you have migrated the configurations for Apache Accumulo from the source cluster to the OpDB service on the destination cluster.
3. Migrate Accumulo data.
 - a) Export each table on the source cluster.
 - b) Collect the exported table data for transfer.
 - c) Move the collected data into the HDFS storage on the destination cluster.
 - d) Install and configure the Accumulo on CDP service on the destination cluster.
 - e) Import each table into the destination cluster.

For more information, see [Exporting Tables](#) in Apache Accumulo documentation.

Post-migration steps

After you configure your destination cluster and migrate your data and workloads, there may be additional steps required.

Sidecar migration for Cloudera Flow Management

You can migrate Apache NiFi and Apache NiFi Registry to CDP Private Cloud Base.

Before you begin

The migration instructions apply to Cloudera Flow Management (CFM) version 1.1.0 (standalone, not managed by Cloudera Manager).

Procedure

Follow the applicable data migration path in [Upgrade and migration paths](#).

Testing applications

After your data, configuration, and workloads are migrated to the destination cluster, you should thoroughly test your applications before moving the cluster to production.

Testing your applications has several primary goals:

- Ensure that your application code has been properly updated to use any new APIs in the services.
- Ensure that you compile your applications using the [Maven artifacts](#) for the new version of CDP Private Cloud Base.
- Ensure that your applications and queries return the expected results.
- Ensure that your applications perform as well or better than how they performed on the source cluster. You can use [Cloudera Workload XM](#) to analyze the performance on the source cluster, and then compare those results with a similar analysis on the destination cluster.

Once testing is complete and the cluster is performing as expected, you can move the destination cluster into production and decommission the source cluster.