

COD Data Migration

HDP 2.6.5 to COD (HBase snapshot)

Version 1.0

*Document Prepared by:*

Ramprasad Ohnu, Solutions Consultant

Cloudera Professional Services

|  |  |  |  |
| --- | --- | --- | --- |
| **Revision History** | | | |
| **Version** | **Author** | **Description** | **Date** |
| 1.0 | Ramprasad Ohnu (rganeshbabu@cloudera.com) | U version | 09/23/2020 |
|  |  |  |  |

**Important Notice**

© 2010-2020 Cloudera, Inc. All rights reserved.

Cloudera, the Cloudera logo, and any other product or service names or slogans contained in this document, except as otherwise disclaimed, are trademarks of Cloudera and its suppliers or licensors, and may not be copied, imitated or used, in whole or in part, without the prior written permission of Cloudera or the applicable trademark holder.

Hadoop and the Hadoop elephant logo are trademarks of the Apache Software Foundation. All other trademarks, registered trademarks, product names and company names or logos mentioned in this document are the property of their respective owners to any products, services, processes or other information, by trade name, trademark, manufacturer, supplier or otherwise does not constitute or imply endorsement, sponsorship or recommendation thereof by us.

Complying with all applicable copyright Azure is the responsibility of the user. Without limiting the rights under copyright, no part of this document may be reproduced, stored in or introduced into a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording, or otherwise), or for any purpose, without the express written permission of Cloudera.

Cloudera may have patents, patent applications, trademarks, copyrights, or other intellectual property rights covering subject matter in this document. Except as expressly provided in any written license agreement from Cloudera, the furnishing of this document does not give you any license to these patents, trademarks copyrights, or other intellectual property.

The information in this document is subject to change without notice. Cloudera shall not be liable for any damages resulting from technical errors or omissions which may be present in this document, or from use of this document.

**Cloudera, Inc.**

395 Page Mill Road Palo Alto, CA 94306  [info@cloudera.com](mailto:info@cloudera.com)   US: 1-888-789-1488  Intl: 1-650-362-0488  [www.cloudera.com](http://www.cloudera.com)

**Release Information**Version: 1.0 Date: 9/23/2020

[**Overview**](#_heading=h.7av0za9zdia8) **5**

[**Pre-requisites**](#_heading=h.qv6fqy73qtan) **5**

[**Configure an edge node on Azure**](#_heading=h.gjdgxs) **5**

[**Configure network line-of-sight**](#_heading=h.fsxxjf1b0z01) **5**

[**Re-use the subnets created for CDP**](#_heading=h.1vclj1l7j6y0) **6**

[**Direct Connect**](#_heading=h.yw31euh35lr7) **6**

[**Virtual Private Network**](#_heading=h.g569vgwb44v2) **7**

[**Configure DNS**](#_heading=h.8i1q2183h7yu) **7**

[**Configure the nameserver**](#_heading=h.koi49m7wwvl9) **8**

[**Configure Kerberos**](#_heading=h.vl2kefe9kwxz) **9**

[**Download Client Configuration**](#_heading=h.21nitfoxxeaq) **10**

[**Using the HBase client tarball**](#_heading=h.long3txowoq2) **10**

[**HDP: Export snapshot to ADLS**](#_heading=h.30j0zll) **11**

[**HDFS Client Configuration**](#_heading=h.2it49juer4cs) **11**

[**ADLS Connectivity check between HDP and COD**](#_heading=h.ykuh9tjzranp) **11**

[**HBase Snapshots**](#_heading=h.j4pomstbjo3u) **12**

[**HBase Configuration**](#_heading=h.u20k5oo33m5w) **12**

[**Export snapshot to ADLS**](#_heading=h.4ebeesual2mi) **13**

[**COD: Create a table from a snapshot in ADLS**](#_heading=h.yy9etjy05cye) **13**

[**First verify that the COD instance sees the snapshot**](#_heading=h.hrw6j96plkuk) **13**

[**Data Validation**](#_heading=h.rka9jeufeqm7) **14**

## 

# Overview

This data migration document covers one possible approach to loading an existing on premise data set into COD. It presumes the source environment is running HDP 2.6.5 and that the destination is a Technical Preview COD database in a CDP environment with CDP Runtime 7.2.1. It only works if the HDP tables have never used the “prefix encoding” option for Apache HBase.

# Pre-requisites

1. Priorly, CDP Environment should be created
2. a CDP workload username with Environment Admin access
3. [Configure an edge node on Azure](#_heading=h.gjdgxs)
4. Needed HBase Client tar for COD to access hbase shell[[1]](#footnote-1)
5. HDFS access to ADLS Refer: [HDP: Export snapshot to ADLS](#_heading=h.30j0zll)
6. Specific location to store snapshots for COD. Navigate to COD Datahub → Cloud Storage → HBase root location
7. Verify ADLS connectivity between HDP 2.6.5 and COD to copy snapshot

# Configure an edge node on Azure

Prerequisites

* Permission to log into the Azure account configured with CDP
* Permission to launch Virtual Machines in your Azure account
* Basic understanding of the Virtual Machines and subnets created for use with CDP
* A CDP environment and a COD database in the environment in which you want this edge node to communicate
* A JSON processing tool such as jq (optional)

To properly access your COD, you must configure these on your edge node:

* Network line-of-sight
* DNS entries (*/etc/resolv.conf*)
* Kerberos KDC configuration (*/etc/krb5.conf*)

## Configure network line-of-sight

An edge node must be able to access the HBase services running inside of your COD. These services are bound to the private network interfaces. Therefore, your edge node must be able to resolve the private IP addresses in your CDP VNet. You can accomplish this by using one of the following methods; however, this is not an exhaustive list. See the Azure documentation for more information about each of these methods.

### Re-use the subnets created for CDP

As a part of the initial CDP setup, you have already created three subnets or CDP would have automatically created this for you. You can use the standard computer service in the Azure Marketplace to launch a Virtual Machine on the same resource group and subnet as your COD. You can choose a Linux distribution, the size of the instance, and the authentication based on your requirements and security policies. Ensure that you allocate enough resources for the edge node based on your use case and requirements.

**Important**: Launching a new Virtual Machine on the same subnet as COD reduces the number of IP addresses that CDP can use to automatically provision new resources for you.

To reuse the subnets:

1. Create a Virtual Machine from the Azure Marketplace. For example, CentOS, the standard instance with an image of your choice.
   1. **Note**: You can use a public key or password authentication based on your security requirements. Spot instances are not recommended for the edge node.
2. Set your inbound port rules to allow access to this edge node by both the COD as well as your client applications.
3. Click **Next : > Disks.**
4. Select a storage type of your choice.
5. Click **Next: Networking >**.
6. Select your CDP VNet.
7. Select a subnet from the VNet
8. If you have created both public and private subnets, choose the public subnet.
9. **Public IP** that is accessible from the Internet.
10. Review your settings and then click **Review + Create**.

For more information about creating an Azure Virtual Machine, see [Azure Virtual Machines](https://docs.microsoft.com/en-us/azure/virtual-machines/).

Verify that the COD edge node can access Hbase Master and its region server as well as to Zookeeper via the security group to the COD cluster.

* 1. You will need to verify port settings for at least the following
     1. HBase Master and all HBase Region Servers
     2. Zookeeper Servers

https://docs.cloudera.com/cloudera-manager/7.1.1/installation/topics/cdpdc-ports-used-by-runtime.html

### Direct Connect

To enable computers in your corporate network to communicate with the CDP VNet, you can use Azure ExpressRoute. Azure ExpressRoute enables hardware which you control to communicate with resources on Azure. For more information about Direct Connect, see [Azure ExpressRoute Overview: Connect over a private connection](https://docs.microsoft.com/en-us/azure/expressroute/expressroute-introduction).

### Virtual Private Network

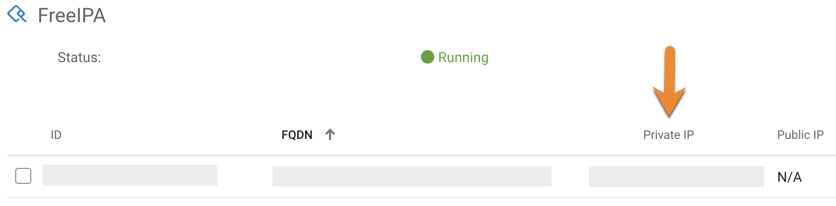
If you are using a corporate VPN network, you can use establish a VPN connection over the Internet to your COD. For more information, see [About Azure VPN Gateway](https://docs.microsoft.com/en-us/azure/vpn-gateway/vpn-gateway-about-vpngateways)**.**

Post-req

After you have created your edge node and can SSH into it, validate that you have a networking line-of-sight correctly configured. You can validate this using the ping command from your edge node to an IP address from within your CDP VNet.

To validate

1. Go to the **CDP Control Plane >  Management Console**.
2. Click **Environments**. And then select your environment from the list of environments.
3. Click **Summary**.
4. Find the FreeIPA section and copy the content in the Private IP field.



1. Open a terminal.
2. Run the command ping <Private IP>

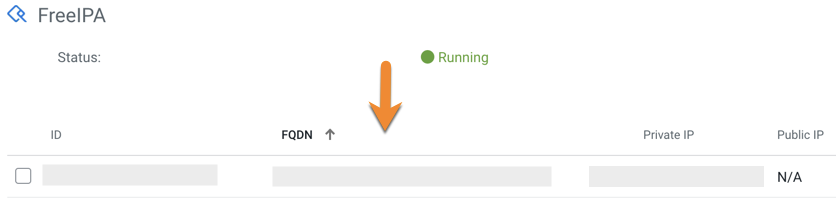
You have successfully configured your edge node if the host responds to the ping request. The ping failing to reply does not necessarily indicate that you have failed to configure the edge node. You may have configured your CDP VNet to have security groups which do not respond to ping, but this indicates that you should take additional steps to validate your line of sight.

## Configure DNS

An edge node must have the ability to resolve the hostnames from your COD. You must obtain the private IP address of the name server and configure it to resolve hostnames from your COD. Each CDP environment acts as its own DNS nameserver. There are multiple ways to obtain the nameserver for your environment.

Obtain the nameserver IP address using the CDP Control Plane

1. Go to the **CDP Control Plane > Management Console**.
2. Click **Environments**. And then select your environment from the list of environments.
3. Click **Summary**.
4. Find the FreeIPA section and copy the content in the Private IP field.



Obtain the nameserver IP address using CDP-CLI

You can also get this IP address using the cdpcli and the jq utility.

Run this command to get the nameserver IP address:

$ cdp environments describe-environment --environment-name <your\_environment> | jq '.environment.freeipa

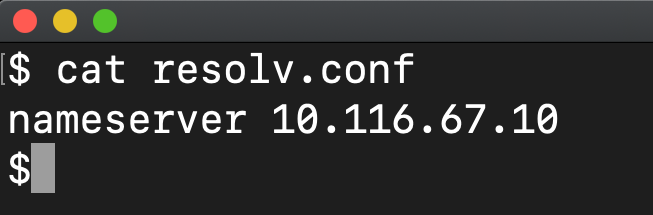
The nameserver IP address is displayed as the output. For example, 10.116.67.10.

### Configure the nameserver

Once you have obtained the private IP address of the nameserver, modify the file */etc/resolv.conf* on your edge node to list this private IP address as a “nameserver”.

$cat resolv.conf

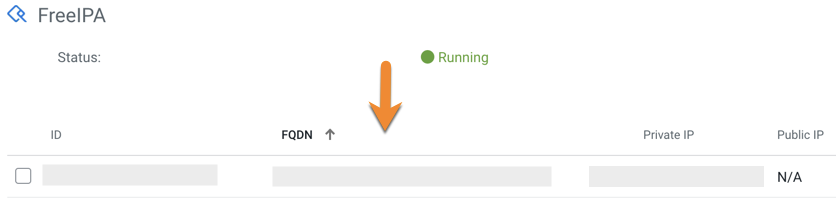
nameserver <nameserver ip address>



Post-req

To validate that you can now resolve DNS names, you can use the command line tool nslookup.

1. Go to the **CDP Control Plane >  Management Console**.
2. Click **Environments**. And then select your environment from the list of environments.
3. Click **Summary**.
4. Find the FreeIPA section and copy the content in the FQDN field.



1. Run this command in your terminal.

$ nslookup <fqdn>

Your DNS is set up correctly If this command returns an address for the name you provided.

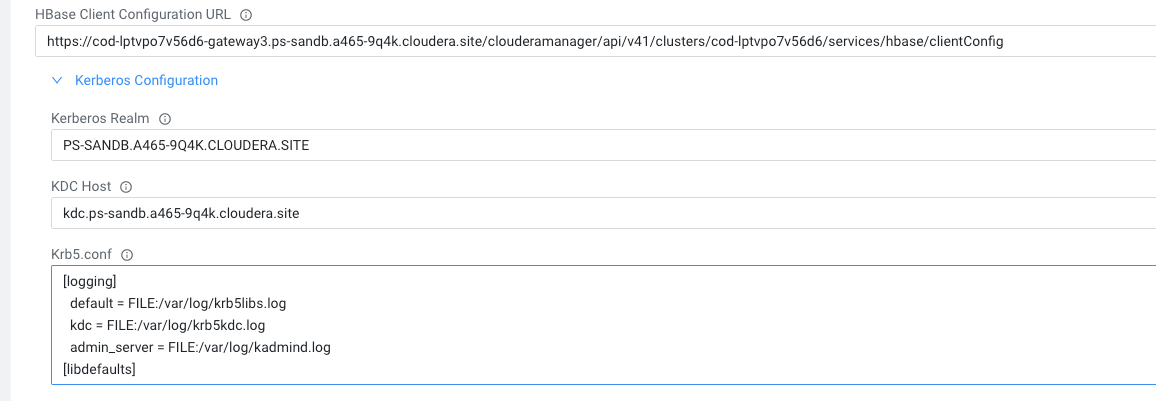
## Configure Kerberos

All CODs are secured with Kerberos-based authentication. This means that only authorized users can connect to your database. All HBase and Phoenix Thick JDBC clients must have a proper Kerberos configuration on the host where they run a client.

The following command provides the necessary Kerberos information as well as a sufficient krb5.conf file encoded with Base64 to use. You can run this command in the terminal to capture the Kerberos configuration.

$ cdp opdb get-client-connectivity --environment-name <env> \  --database-name <name> | jq -r \  ‘.kerberosConfiguration.krb5Conf’ | base64 --decode

or copy the config from COD UI out of the “Krb5.conf” text box:



Copy the output of this command, and place the contents into the file */etc/krb5.conf* on your edge node.

Post-req

To validate that Kerberos is correctly set up, you can use the command kinit to validate that you are able to obtain a Kerberos ticket. Run this command to validate.

$ kinit <cdp\_workload\_name>

Password: <cdp\_workload\_password>

Create Workload password in CDP → User Management

If you successfully authenticate, you will not receive an error and will be able to validate that you have a ticket using the command klist.

## Download Client Configuration

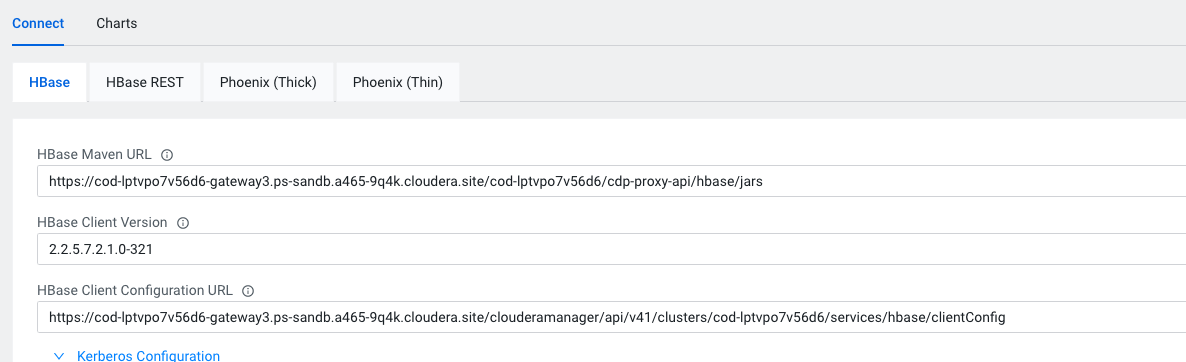
use the clientConfigurationUrl from describe-client-connectivity to obtain the necessary configuration clients to talk to HBase

$ cdp opdb describe-client-connectivity --database-name <database-name>

--environment-name <environment-name>

  | jq '.connectors[] | select(.name == "hbase") | .configuration.clientConfigurationUrl'

We copy the Client configuration URL from COD UI:



$ curl -k -o clientConfig.zip -u '<cdp\_workload\_name>:<cdp\_workload\_password>' <config\_url>  unzip clientConfig.zip to get hbase-conf folder

## Using the HBase client tarball

HBase client libraries will be provided in a common location that the team can all access to run HBase shell commands from edge node.

Note:- CDP 7.2.1 client tarball is an interim solution to access HBase.   Permission:- The CDP workload username you provide must have Environment Admin access to use HBase administrative commands.

1) Copy the tarball to the edge node into the directory you intend to use.

> tar -xf hbase-client-tarball.tar.gz

2) To use Shell functionality: -

 > export HBASE\_CONF\_DIR=<path of COD downloaded hbase config folder>

> kinit <cdp\_workload\_username>.

> klist

> bin/hbase shell

# HDP: Export snapshot to ADLS

## HDFS Client Configuration

Steps to export snapshots from HDFS to ADLS.

In order to access data stored in your Azure blob storage account, you must configure your storage account access key in core-site.xml. The configuration property that you must use is fs.azure.account.key.<account name>.blob.core.windows.net and the value is the access key.

<property>

<name>fs.azure.account.key.testaccount.blob.core.windows.net</name>

<value>TESTACCOUNT-ACCESS-KEY</value>

</property>

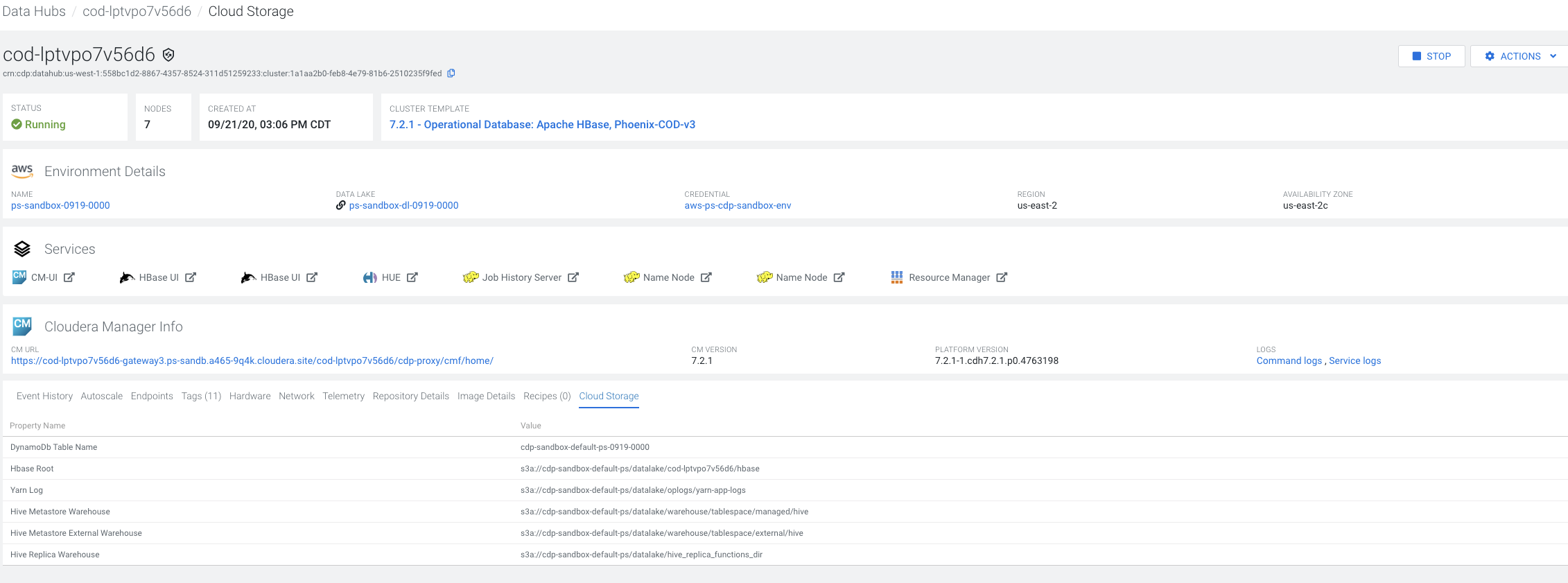
## Refer: <https://docs.cloudera.com/HDPDocuments/HDP2/HDP-2.6.5/bk_cloud-data-access/content/authentication-wasb.html>

## ADLS Connectivity check between HDP and COD

COD databases are created within a CDP environment that includes access to a specific backing ADLS bucket. The steps below will allow us to verify that the existing HDP cluster has the needed configuration settings to write into that bucket when running the snapshot export tool. While the cluster may already have access to *some* ADLS bucket for DR backups, these steps do not assume that the same access will work for getting to the CDP environment’s object storage.

1. Determine the ADLS bucket path used by COD.

Navigate into the corresponding Datahub page for your COD database, and click the "Cloud Storage" tab, look at the value for "HBase root dir"



1. On the HDP cluster, run a listing command on the URL given. It should look like:

hadoop fs -ls abfs://:<storagename>/datalake/<datahub\_name>/hbase

* 1. If the file listing *succeeds*, then you should see several directories for HBase internals and the HDP cluster can read from that path.
  2. *fails*, then the listing may be blank or there may be an error message. There should be a non-zero exit code. You may need to re-run the command, setting the appropriate environment variables to configure an allowed Azure key for the appropriate ADLS bucket.

## HBase Snapshots

### HBase Configuration

The snapshot operation doesn’t involve any data copying. Since we are getting a representative sample of current production data for use in a test system it is safe for us to use the skip\_flush option to also save the IO overhead of getting edits currently in progress flushed to Apache HDFS. If we were e.g. bootstrapping a DR replication peer then we would need to pay this overhead.

For more snapshot options, refer to [HBase Snapshots](http://hbase.apache.org/1.1/book.html#ops.snapshots) section in [HBase reference guide](http://hbase.apache.org/1.1/book.html).

$ hbase shell

hbase> snapshot 'myTable', 'myTableSnapshot-122112' {SKIP\_FLUSH => true}

### Export snapshot to ADLS

|  |
| --- |
| $ sudo -u hbase  $ hbase org.apache.hadoop.hbase.snapshot.ExportSnapshot -Dsnapshot.export.skip.tmp=true -snapshot **<snapshot\_name>** -copy-to wasbs://data@<storage\_account\_name>/<datahub\_name>/hbase |

Verify that snapshot was copied to ADLS.

|  |
| --- |
| $ hbase org.apache.hadoop.hbase.snapshot.SnapshotInfo -snapshot **<snapshot\_name>** -remote-dir wasbs://data@<storage\_account\_name>/<datahub\_name>/hbase |

# COD: Create a table from a snapshot in ADLS

Presuming a snapshot exists within the ADLS bucket backing a COD database, you can use the following examples to get that data into a live table. The instructions assume any table names given do not already exist in the COD database.

### Verify snapshot is available in COD

Using the hbase client installation on the edge node, list snapshots in the shell

$./bin/hbase shell

hbase> list\_snapshots SNAPSHOT TABLE + CREATION TIME

myTableSnapshot-122112 myTable(2020-09-23 22:13:39 UTC)  => ["myTableSnapshot-122112"]

Note that the snapshot name you used above should be what is listed if things were correctly copied into the ADLS bucket backing your COD database. Next use the clone command to create a table in the COD database based on the data in the snapshot. This action does not involve any data copying and should be quick

$./bin/hbase shell

hbase> clone\_snapshot 'myTableSnapshot-122112', 'myNewTestTable'  
 Took <time> seconds

# Data Validation

You should be able to verify the data in this new table with whatever existing validation you have from the edge node provided those processes only rely on HBase client access.   In the absence of an existing process you can use the built-in count command to check for a suitable number of rows.

hbase> count 'myNewTestTable'

1. [↑](#footnote-ref-1)