

Cloudera Runtime 7.0.2

Getting Started with Apache HBase

Date published: 2019-09-23

Date modified:

CLouDERA

<https://docs.cloudera.com/>

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

Operational database cluster.....	4
Before you create an operational database cluster.....	4
Creating an operational database cluster.....	5
Default operational database cluster definition.....	5
Provision an operational database cluster.....	5

Operational database cluster

You can create an HBase cluster along with all the required services such as HDFS and ZooKeeper using the Operational Database Data Hub cluster template in CDP. The operational database cluster uses Amazon S3 as a storage layer for HBase, where HFiles are written to S3, but WALs are written to HDFS.

The Operational Database cluster template in CDP consists of HDFS, HBase, and ZooKeeper. CDP makes use of reusable cluster templates. The cluster template for Operational Database consists of Apache HBase and the underlying HDFS and Apache ZooKeeper services that it needs to operate. Knox is used for the proxying of the different user interfaces.

Each cluster definition must reference a specific cluster template. You can also create and upload your cluster templates. For example, you can Export a cluster template from an existing on-prem HDP cluster and register it in CDP to use it for creating Data Hub clusters.

Related Information

[Cluster Templates](#)

Before you create an operational database cluster

You must create an environment, register the environment, and set your user's password before you can create an operational database cluster.

Before you create an Operational Database cluster as a Data Hub cluster in CDP, you must ensure that you have completed the following steps:

Set up user access and authorization

To provide access to resources such as environments and clusters, you must add users and groups and assign roles and resources to them. Each environment has an IPA instance which acts as a centralized user database. To access non-SSO interfaces, each user must set an IPA password.

For more information, see *Managing user access and authorization*.

Create an environment on Amazon AWS

You require the following to create an environment on Amazon AWS for use with CDP:

- Cloud credential to configure access from CDP to your cloud provider account
- SSH key-pair in CDP. You can generate an access key from the Cloudera Management Console using User Management
- Knowledge of your AWS Region and VPC/Network information

For more information, see *Environment prerequisites on AWS*.

Register your environment in CDP

You must register an environment before you can create a cluster on that environment. The environment will retain all of this information for any clusters created within it and will create a FreeIPA cluster for centralized user management.

You can register an environment from the Cloudera Management Console using Environments > Register Environment.

For more information, see *Register an AWS environment*.



Note: Your OpDB cluster may fail to initialize if you are reusing a cloud storage location that was used by an older OpDB Data Hub cluster. If you want to reuse a cloud storage location, you should first stop HBase using Cloudera Manager before deleting an OpDB Data Hub cluster. For instructions to stop HBase, see [Stop HBase](#). If these steps are not followed, the new OpDB Data Hub cluster may start with regions that are not correctly assigned because of the state of the old OpDB Data Hub cluster, which will cause HBase to fail to fully start, commonly with the error "Master is initializing. By stopping HBase with Cloudera Manager prior to shutting down an older OpDB Data Hub cluster, you can ensure that a new OpDB Data Hub cluster will start without error when using the same cloud storage location.

Related Information

[Managing user access and authorization](#)

[Setting the IPA Password](#)

[AWS account requirements](#)

[Register an AWS environment](#)

Creating an operational database cluster

After you have set up your environment, you can create an operational database as a Data Hub cluster. Data Hub clusters are those clusters that you can create and manage for your specific use case, and it is powered by Cloudera Runtime.

You can create multiple Data Hub clusters using the same environment. For example, you can create an operational database cluster to replicate data from an on-prem cluster using HBase replication, and using the same environment you can create an operational database cluster that can be used for storing new data.

Default operational database cluster definition

When you create an operational database cluster, you can use the Operational Database template and the default cluster definition. The cluster can be scaled up or down later, but the underlying cluster definition is unchanged to support the master services and the minimum required configuration. The cluster definition primarily defines cloud provider settings.

If you are using Amazon AWS as your public cloud provider, your HBase cluster has the following cluster definition:

- One gateway node that hosts the Knox gateway and Cloudera Manager, and a ZooKeeper Server.
- One node running the HDFS NameNode, and a ZooKeeper Server.
- One node running the HBase Master, and a ZooKeeper Server.
- At least three worker nodes. Each worker node has a RegionServer and a DataNode.



Note:

- You cannot access HBase, HDFS, and ZooKeeper from outside your virtual private cloud (VPC), because there is no network connectivity from outside the VPC. You can SSH to a node in the cluster that you started, or to another node which is within the VPC that is provided to you. The node provided to you should have the HBase client installation and the required HBase configuration files.
- An HDFS instance stores HBase Write Ahead Logs (WALs) because Amazon S3 does not provide the durability guarantees that HBase requires for WALs.

Provision an operational database cluster

You can provision an operational database cluster that has the default cluster definition or you can define the custom number of worker nodes that you need.

About this task

To provision a cluster:

Procedure

1. From the Cloudera Management Console, click Environments.
2. Search for your environment using the search box, and click on the environment name. Your environment clusters page appears.
3. Click Create Data Hub Cluster.
The Provision Data Hub Cluster page appears.
4. From the Selected Environment list, ensure that you have selected the environment in which you want to create the cluster.
5. Choose the way you want to provision your cluster.
 - Cluster Definition: Select the Cluster Definition option to create your cluster quickly by using one of the prescriptive cluster definitions included by default or one of your previously created custom cluster definitions.
 - Custom: Select the Custom option to install selected services as part of the cluster. The list of services is determined by the selected Cluster Template.
6. If you selected Cluster Definition:
 - a) From the Cluster Definition list, select Operational Database .
 - b) In the Cluster Name field, provide a uniquely identifiable cluster name.or
7. If you selected Custom:
 - a) From the Platform Version list, select the CDH version of your choice.
 - b) From the Cluster Template list, select a cluster template of your choice.
 - c) In the Cluster Name field, provide a uniquely identifiable cluster name.
8. Click Tags to add tags to help you find your cluster-related resources in your cloud provider account.
9. Select Advanced Options to configure various options such as Image Catalog, Network And Availability, Hardware And Storage, Cloud Storage, and Cluster Extensions.
10. Set the number of RegionServers by editing the Instance Count in Advanced Options Hardware And Storage .
11. If you want to change the Instance Type of worker nodes to suit your requirements using Advanced Options Hardware And Storage .



Note: The DataNodes and RegionServers are co-located.

12. Click Provision Cluster.

After the cluster is successfully provisioned, you can access the Cloudera Manager user interface from the Cluster Information section.