

# Cloudera Administration



## **Important Notice**

(c) 2010-2015 Cloudera, Inc. All rights reserved.

Cloudera, the Cloudera logo, Cloudera Impala, and any other product or service names or slogans contained in this document 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. Reference 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 laws 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. For information about patents covering Cloudera products, see <http://tiny.cloudera.com/patents>.

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.**  
**1001 Page Mill Road Bldg 2**  
**Palo Alto, CA 94304**  
**[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: 5.3.x  
Date: June 24, 2015

# Table of Contents

<b>About Cloudera Administration.....</b>	<b>6</b>
---	----------

<b>Configuring CDH and Managed Services.....</b>	<b>7</b>
--	----------

Configuring CDH and Managed Services Using Cloudera Manager.....	7
<i>Configuration Overview.....</i>	<i>7</i>
<i>Managing Clusters.....</i>	<i>25</i>
<i>Managing Services.....</i>	<i>29</i>
<i>Managing Roles.....</i>	<i>38</i>
<i>Managing Individual Services.....</i>	<i>43</i>
<i>Managing Hosts.....</i>	<i>96</i>
Configuring and Using CDH from the Command Line.....	109
<i>Starting CDH Services.....</i>	<i>110</i>
<i>Stopping Services.....</i>	<i>115</i>
<i>Migrating Data between a CDH 4 and CDH 5 Cluster.....</i>	<i>118</i>
<i>Configuring HDFS.....</i>	<i>121</i>
<i>Configuring and Using HBase.....</i>	<i>133</i>
<i>Configuring Hive.....</i>	<i>161</i>
<i>Configuring Impala.....</i>	<i>162</i>

<b>Resource Management.....</b>	<b>171</b>
---------------------------------	------------

Schedulers.....	171
Cloudera Manager Resource Management Features.....	171
Managing Resources with Cloudera Manager.....	173
<i>Linux Control Groups.....</i>	<i>173</i>
<i>Static Service Pools.....</i>	<i>176</i>
<i>Dynamic Resource Pools.....</i>	<i>177</i>
<i>Managing Impala Admission Control.....</i>	<i>186</i>
<i>Managing the Impala Llama ApplicationMaster.....</i>	<i>186</i>
Impala Resource Management.....	188
<i>Admission Control and Query Queuing.....</i>	<i>188</i>
<i>Integrated Resource Management with YARN.....</i>	<i>196</i>

<b>Performance Management.....</b>	<b>200</b>
------------------------------------	------------

Improving Performance.....	200
Configuring Short-Circuit Reads.....	203
<i>Configuring Short-Circuit Reads Using Cloudera Manager.....</i>	<i>203</i>

<i>Configuring Short-Circuit Reads Using the Command Line</i> .....	203
Choosing a Data Compression Format.....	204
Tuning the Solr Server.....	205
<i>Tuning to Complete During Setup</i> .....	205
<i>General Tuning</i> .....	206
<i>Other Resources</i> .....	209
<b>High Availability</b> .....	<b>211</b>
HDFS High Availability.....	211
<i>Introduction to HDFS High Availability</i> .....	211
<i>Configuring Hardware for HDFS HA</i> .....	213
<i>Enabling HDFS HA</i> .....	213
<i>Disabling and Redeploying HDFS HA</i> .....	227
<i>Configuring Other CDH Components to Use HDFS HA</i> .....	231
<i>Administering an HDFS High Availability Cluster</i> .....	233
MapReduce (MRv1) and YARN (MRv2) High Availability.....	237
<i>YARN (MRv2) ResourceManager High Availability</i> .....	237
<i>Work Preserving Recovery for YARN Components</i> .....	244
<i>MapReduce (MRv1) JobTracker High Availability</i> .....	245
High Availability for Other CDH Components.....	258
<i>HBase High Availability</i> .....	258
<i>Hive Metastore High Availability</i> .....	258
<i>Llama High Availability</i> .....	260
<i>Oozie High Availability</i> .....	262
<i>Search High Availability</i> .....	263
<b>Backup and Disaster Recovery</b> .....	<b>266</b>
Backup and Disaster Recovery Overview.....	266
Data Replication.....	267
<i>Designating a Replication Source</i> .....	268
<i>HBase Replication</i> .....	269
<i>HDFS Replication</i> .....	274
<i>Hive Replication</i> .....	276
<i>Impala Metadata Replication</i> .....	279
<i>Enabling Replication Between Clusters in Different Kerberos Realms</i> .....	279
Snapshots.....	280
<i>Cloudera Manager Snapshot Policies</i> .....	280
<i>Managing HBase Snapshots</i> .....	282
<i>Managing HDFS Snapshots</i> .....	292
<b>Cloudera Manager Administration</b> .....	<b>295</b>
Managing the Cloudera Manager Server and Agents.....	295

<i>Starting, Stopping, and Restarting the Cloudera Manager Server.....</i>	<i>295</i>
<i>Configuring Cloudera Manager Server Ports.....</i>	<i>295</i>
<i>Moving the Cloudera Manager Server to a New Host.....</i>	<i>295</i>
<i>Starting, Stopping, and Restarting Cloudera Manager Agents.....</i>	<i>296</i>
<i>Configuring Cloudera Manager Agents.....</i>	<i>297</i>
<i>Managing Cloudera Manager Server and Agent Logs.....</i>	<i>300</i>
<i>Changing Hostnames.....</i>	<i>301</i>
<i>Configuring Network Settings.....</i>	<i>304</i>
<i>Managing Alerts.....</i>	<i>304</i>
<i>Managing Licenses.....</i>	<i>306</i>
<i>Sending Usage and Diagnostic Data to Cloudera.....</i>	<i>311</i>
<i>Exporting and Importing Cloudera Manager Configuration.....</i>	<i>314</i>
<i>Other Cloudera Manager Tasks and Settings.....</i>	<i>314</i>
<i>Displaying the Cloudera Manager Server Version and Server Time.....</i>	<i>315</i>
<i>Displaying Cloudera Manager Documentation.....</i>	<i>315</i>
Cloudera Management Service.....	316

## **Cloudera Navigator Administration.....321**

Cloudera Navigator Audit Server.....	321
Cloudera Navigator Metadata Server.....	324
Displaying the Cloudera Navigator Version.....	330
Displaying Cloudera Navigator Documentation.....	330

## About Cloudera Administration

This guide describes how to configure and administer a Cloudera deployment. Administrators manage resources, availability, and backup and recovery configurations. In addition, this guide shows how to implement high availability, and discusses integration.

# Configuring CDH and Managed Services

## Configuring CDH and Managed Services Using Cloudera Manager

You configure CDH and managed services using the [Cloudera Manager Admin Console](#) and [Cloudera Manager API](#).

The following sections focus on the Cloudera Manager Admin Console.

### Configuration Overview

When Cloudera Manager configures a service, it allocates one or more functions (called **roles** in Cloudera Manager) that are required for that service to the hosts in your cluster. The role determines which service daemons run on a given host. For example, when Cloudera Manager configures an HDFS service instance it configures one host to run the NameNode role, another host to run as the Secondary NameNode role, another host to run the Balancer role, and some or all of the remaining hosts as to run DataNode roles.

A **role group** is a set of configuration properties for a role type, as well as a list of role instances associated with that group. Cloudera Manager automatically creates a default role group named **Role Type Default Group** for each role type.

When you run the installation or upgrade wizard, Cloudera Manager automatically creates the appropriate configurations for the default role groups it adds. It may also create additional role groups for a given role type, if necessary. For example, if you have a DataNode role on the same host as the NameNode, it may require a slightly different configuration than DataNode roles running on other hosts. Therefore, Cloudera Manager will create a separate role group for the DataNode role that is running on the NameNode host, and use the default DataNode configuration for the DataNode roles running on other hosts.

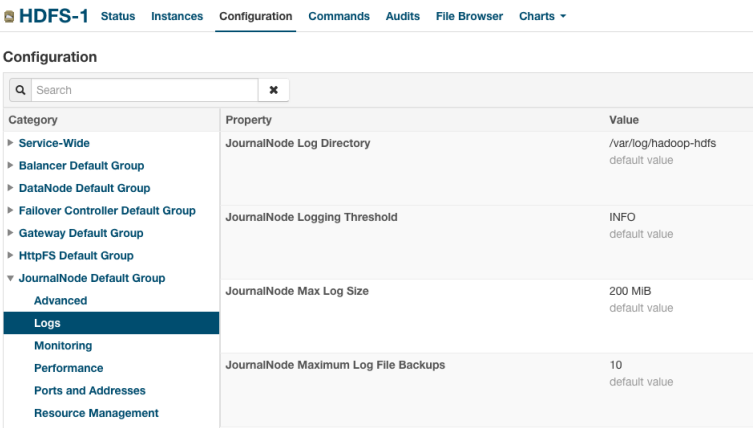
Cloudera Manager wizards [autoconfigure](#) role group properties based on the resources available on the hosts. For properties that are not dependent on host resources, defaults typically align with whatever CDH uses by default for that configuration. Cloudera Manager generally only deviates when the CDH default is not a recommended configuration. Sometimes default values are illegal, such as for data directories (the default is no data directories). The complete catalog of properties and their default values are documented in [Cloudera Manager Configuration Properties](#).

After running the First Run installation wizard, you can use Cloudera Manager to reconfigure the existing services, and add and configure additional hosts and services.

The Cloudera Manager configuration screens offer two layout options: classic and new. The classic layout is the default; however, on each configuration page you can easily switch between the classic and new layouts using the **Switch to XXX layout** link at the top right of the page. All the configuration procedures described in the Cloudera Manager documentation assume the classic layout.

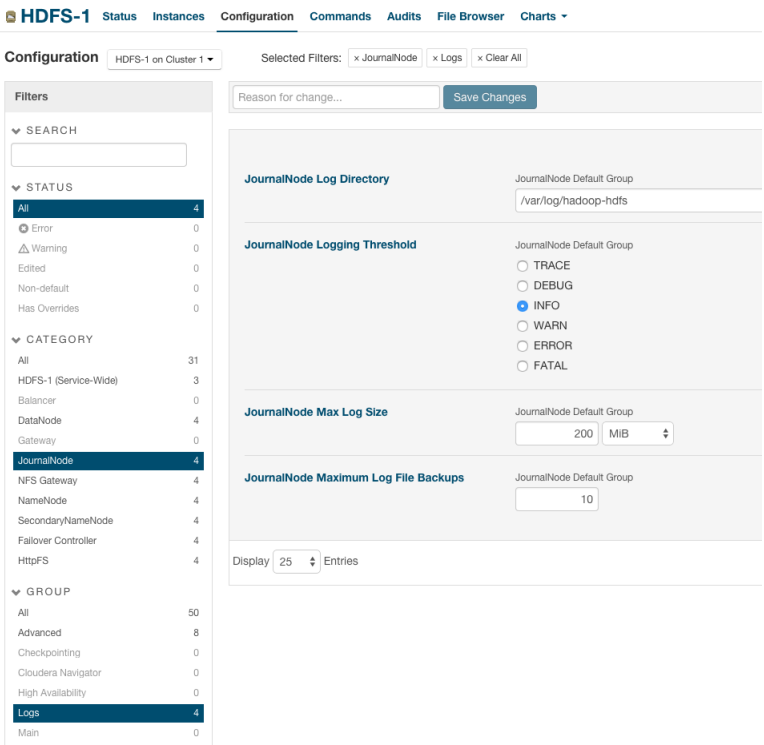
- **classic** - pages are organized by role group and categories within the role group. For example, to display the JournalNode maximum log size property (JournalNode Max Log Size), select **JournalNode Default Group > Logs**.

# Configuring CDH and Managed Services



When a configuration property has been set to a value different from the default, a **Reset to the default value** link displays.

- **new** – pages contain controls that allow you filter configuration properties based on configuration status, category, and group. For example, to display the JournalNode maximum log size property (JournalNode Max Log Size), click the **CATEGORY > JournalNode** and **GROUP > Logs** filters:



When a configuration property has been set to a value different from the default, a reset to default value icon displays.

There is no mechanism for resetting to an [autoconfigured](#) value. However, you can use the configuration [history and rollback feature](#) to revert any configuration changes.




## Modifying Configuration Properties

Required Role: **Configurator** **Cluster Administrator** **Full Administrator**



When a service is added to Cloudera Manager, either through the installation or upgrade wizard or with the Add Services workflow, Cloudera Manager automatically sets the configuration properties, based on the needs of the service and characteristics of the cluster in which it will run. These configuration properties include both service-wide configuration properties, as well as specific properties for each role type associated with the service, managed through role groups. A **role group** is a set of configuration properties for a role type, as well as a list of role instances associated with that group. Cloudera Manager automatically creates a default role group named **Role Type Default Group** for each role type. See [Role Groups](#) on page 42.

### Changing the Configuration of a Service or Role Instance

1. Go to the service status page.
2. Click the **Configuration** tab.
3. Under the appropriate role group, select the category for the properties you want to change.
4. To search for a text string (such as "snippet"), in a property, value, or description, enter the text string in the **Search** box at the top of the category list.
5. Moving the cursor over the value cell highlights the cell; click anywhere in the highlighted area to enable editing of the value. Then type the new value in the field provided (or check or uncheck the box, as appropriate).
  - To facilitate entering some types of values, you can specify not only the value, but also the units that apply to the value. For example, to enter a setting that specifies bytes per second, you can choose to enter the value in bytes (B), KiBs, MiBs, or GiBs—selected from a drop-down menu that appears when you edit the value.
  - If the property allows a list of values, click the **+** icon to the right of the edit field to add an additional field. An example of this is the HDFS DataNode Data Directory property, which can have a comma-delimited list of directories as its value. To remove an item from such a list, click the **−** icon to the right of the field you want to remove.
6. Click **Save Changes** to commit the changes. You can add a note that will be included with the change in the Configuration History. This will change the setting for the role group, and will apply to all role instances associated with that role group. Depending on the change you made, you may need to restart the service or roles associated with the configuration you just changed. Or, you may need to redeploy your client configuration for the service. You should see a message to that effect at the top of the Configuration page, and services will display an outdated configuration  (Restart Needed),  (Refresh Needed), or outdated client configuration  indicator. Click the indicator to display the [Stale Configurations](#) on page 21 page.

### Validation of Configuration Properties

Cloudera Manager validates the values you specify for configuration properties. If you specify a value that is outside the recommended range of values or is invalid, Cloudera Manager displays a warning at the top of the **Configuration** tab and in the text box after you click **Save Changes**. The warning is yellow if the value is outside the recommended range of values and red if the value is invalid.

### Overriding Configuration Properties

For role types that allow multiple instances, each role instance inherits its configuration properties from its associated role group. While role groups provide a convenient way to provide alternate configuration properties for selected groups of role instances, there may be situations where you want to make a one-off configuration change—for example when a host has malfunctioned and you want to temporarily reconfigure it. In this case, you can override configuration properties for a specific role instance:

1. Go to the **Status** page for the service whose role you want to change.
2. Click the **Instances** tab.
3. Click the role instance you want to change.
4. Click the **Configuration** tab.
5. Change the configuration values as appropriate.
6. Save your changes.


You will most likely need to restart your service or role to have your configuration changes take effect.

Viewing and Editing Overridden Configuration Properties

To see a list of all role instances that have an override value for a particular configuration setting, navigate to the entry for the configuration setting in the Status page, expand the **Overridden by *n* instance(s)** link in the value cell for the overridden value.

▼ Overridden by 1 instance(s)

5.0 GiB (DATANODE tcdn5-2.ent.cloudera.com)

 [Edit Overrides](#)

To view the override values, and change them if appropriate, click the **Edit Overrides** link. This opens the **Edit Overrides** page, and lists the role instances that have override properties for the selected configuration setting.

**Edit Overrides: DataNode Default Group - Reserved Space for Non DFS Use**

Reserved space in bytes per volume for non Distributed File System (DFS) use.

Change value of selected instances to: 

Inherited Value ▼

Apply

<input type="checkbox"/>	Role Name	Value
		Overrides Only ▼
<input type="checkbox"/>	datanode (tcdn5-2)	5 GiB

On the **Edit Overrides** page, you can do any of the following:

- View the list of role instances that have overridden the value specified in the role group. Use the selections on the drop-down menu below the **Value** column header to view a list of instances that use the inherited value, instances that use an override value, or all instances. This view is especially useful for finding inconsistent properties in a cluster. You can also use the **Host** and **Rack** text boxes to filter the list.
- Change the override value for the role instances to the inherited value from the associated role group. To do so, select the role instances you want to change, choose **Inherited Value** from the drop-down menu next to **Change value of selected instances to** and click **Apply**.
- Change the override value for the role instances to a different value. To do so, select the role instances you want to change, choose **Other** from the drop-down menu next to **Change value of selected instances to**. Enter the new value in the text box and then click **Apply**.

Resetting Configuration Properties to the Default Value

To reset a property back to its default value, click the **Reset to the default value** link below the text box in the value cell. The default value is inserted and both the text box and the Reset link disappear. Explicitly setting a configuration to the same value as its default (inherited value) has the same effect as using the **Reset to the default value** link.

There is no mechanism for resetting to an [autoconfigured](#) value. However, you can use the configuration [history and rollback feature](#) to revert any configuration changes.

Restarting Services and Instances after Configuration Changes

If you change the configuration properties after you start a service or instance, you may need to restart the service or instance to have the configuration properties become active. If you change configuration properties at the service level that affect a particular role only (such as all DataNodes but not the NameNodes), you can

restart only that role; you do not need to restart the entire service. If you changed the configuration for a particular role instance (such as one of four DataNodes), you may need to restart only that instance.

1. Follow the instructions in [Restarting a Service](#) on page 34 or [Starting, Stopping, and Restarting Role Instances](#) on page 40.
2. If you see a **Finished** status, the service or role instances have restarted.
3. Navigate to the Home page. The service should show a Status of **Started** for all instances and a health status of **Good**.

For further information, see [Stale Configurations](#) on page 21.

## Autoconfiguration

Cloudera Manager provides several interactive wizards to automate common workflows:

- Installation - used to bootstrap a Cloudera Manager deployment
- Add Cluster - used when adding a new cluster
- Add Service - used when adding a new service
- Upgrade - used when upgrading to a new version of Cloudera Manager
- Static Service Pools - used when configuring static service pools
- Import MapReduce - used when migrating from MapReduce to YARN

In some of these wizards, Cloudera Manager uses a set of rules to automatically configure certain settings to best suit the characteristics of the deployment. For example, the number of hosts in the deployment drives the memory requirements for certain monitoring daemons: the more hosts, the more memory is needed. Additionally, wizards that are tasked with creating new roles will use a similar set of rules to determine an ideal host placement for those roles.

## Scope

The following table shows, for each wizard, the scope of entities it affects during autoconfiguration and role-host placement.

Wizard	Autoconfiguration Scope	Role-Host Placement Scope
Installation	New cluster, Cloudera Management Service	New cluster, Cloudera Management Service
Add Cluster	New cluster	New cluster
Add Service	New service	New service
Upgrade	Cloudera Management Service	Cloudera Management Service
Static Service Pools	Existing cluster	N/A
Import MapReduce	Existing YARN service	N/A

Certain autoconfiguration rules are unscoped, that is, they configure settings belonging to entities that aren't necessarily the entities under the wizard's scope. These exceptions are explicitly listed.

## Autoconfiguration

Cloudera Manager employs several different rules to drive automatic configuration, with some variation from wizard to wizard. These rules range from the simple to the complex.

## Configuration Scope

One of the points of complexity in autoconfiguration is configuration scope. The configuration hierarchy as it applies to services is as follows: configurations may be modified at the service level (affecting every role in the service), [role group](#) level (affecting every role instance in the group), or role level (affecting one role instance). A configuration found in a lower level takes precedence over a configuration found in a higher level.

## Configuring CDH and Managed Services

With the exception of the Static Service Pools, and the Import MapReduce wizard, all Cloudera Manager wizards follow a basic pattern:

1. Every role in scope is moved into its own, new, role group.
2. This role group is the receptacle for the role's "idealized" configuration. Much of this configuration is driven by properties of the role's host, which can vary from role to role.
3. Once autoconfiguration is complete, new role groups with common configurations are merged.
4. The end result is a smaller set of role groups, each with an "idealized" configuration for some subset of the roles in scope. A subset can have any number of roles; perhaps all of them, perhaps just one, and so on.

The Static Service Pools and Import MapReduce wizards configure role groups directly and do not perform any merging.

### Static Service Pools

Certain rules are only invoked in the context of the Static Service Pools wizard. Additionally, the wizard autoconfigures cgroup settings for certain kinds of roles:

- HDFS DataNodes
- HBase RegionServers
- MapReduce TaskTrackers
- YARN NodeManagers
- Impala Daemons
- Solr Servers
- Spark Workers
- Accumulo Tablet Servers
- Add-on services

### YARN

`yarn.nodemanager.resource.cpu-vcores` - For each NodeManager role group, set to ((number of cores, including hyperthreads, on one NodeManager member's host) \* (service percentage chosen in wizard)).

### All Services

Cgroup `cpu.shares` - For each role group that supports `cpu.shares`, set to  $\max(20, (\text{service percentage chosen in wizard}) * 20)$ .

Cgroup `blkio.weight` - For each role group that supports `blkio.weight`, set to  $\max(100, (\text{service percentage chosen in wizard}) * 10)$ .

### Data Directories

Several autoconfiguration rules work with data directories, and there's a common sub-rule used by all such rules to determine, out of all the mountpoints present on a host, which are appropriate for data. The subrule works as follows:

- The initial set of mountpoints for a host includes all those that are disk-backed. Network-backed mountpoints are excluded.
- Mountpoints beginning with `/boot`, `/cdrom`, `/usr`, `/tmp`, `/home`, or `/dev` are excluded.
- Mountpoints beginning with `/media` are excluded, unless the backing device's name contains `/xvd` somewhere in it.
- Mountpoints beginning with `/var` are excluded, unless they are `/var` or `/var/lib`.
- The largest mount point (in terms of total space, not available space) is determined.
- Other mountpoints with less than 1% total space of the largest are excluded.
- Mountpoints beginning with `/var` or equal to `/` are excluded unless they're the largest mount point.
- Remaining mountpoints are sorted lexicographically and retained for future use.

## Memory

The rules used to autoconfigure memory reservations are perhaps the most complicated rules employed by Cloudera Manager. When configuring memory, Cloudera Manager must take into consideration which roles are likely to enjoy more memory, and must not over commit hosts if at all possible. To that end, it needs to consider each host as an entire unit, partitioning its available RAM into segments, one segment for each role. To make matters worse, some roles have more than one memory segment. For example, a Solr server has two memory segments: a JVM heap used for most memory allocation, and a JVM direct memory pool used for HDFS block caching. Here is the overall flow during memory autoconfiguration:

1. The set of participants includes every host under scope as well as every {role, memory segment} pair on those hosts. Some roles are under scope while others are not.
2. For each {role, segment} pair where the role is under scope, a rule is run to determine four different values for that pair:
  - Minimum memory configuration. Cloudera Manager must satisfy this minimum, possibly over-committing the host if necessary.
  - Minimum memory consumption. Like the above, but possibly scaled to account for inherent overhead. For example, JVM memory values are multiplied by 1.3 to arrive at their consumption value.
  - Ideal memory configuration. If RAM permits, Cloudera Manager will provide the pair with all of this memory.
  - Ideal memory consumption. Like the above, but scaled if necessary.
3. For each {role, segment} pair where the role is not under scope, a rule is run to determine that pair's existing memory consumption. Cloudera Manager will not configure this segment but will take it into consideration by setting the pair's "minimum" and "ideal" to the memory consumption value.
4. For each host, the following steps are taken:
  - a. 20% of the host's available RAM is subtracted and reserved for the OS.
  - b.  $\text{sum}(\text{minimum\_consumption})$  and  $\text{sum}(\text{ideal\_consumption})$  are calculated.
  - c. An "availability ratio" is built by comparing the two sums against the host's available RAM.
    - a. If  $\text{RAM} < \text{sum}(\text{minimum})$  ratio = 0
    - b. If  $\text{RAM} \geq \text{sum}(\text{ideal})$  ratio = 1
    - c. Otherwise, ratio is computed via:  $(\text{RAM} - \text{sum}(\text{minimum})) / (\text{sum}(\text{ideal}) - \text{sum}(\text{minimum}))$
5. For each {role, segment} pair where the role is under scope, the segment is configured to be  $(\text{minimum} + ((\text{ideal} - \text{minimum}) * (\text{host availability ratio})))$ . The value is rounded down to the nearest megabyte.
6. The {role, segment} pair is set with the value from the previous step. In the Static Service Pools wizard, the role group is set just once (as opposed to each role).
7. Custom post-configuration rules are run.

Customization rules are applied in steps 2, 3 and 7. In step 2, there's a generic rule for most cases, as well as a series of custom rules for certain {role, segment} pairs. Likewise, there's a generic rule to calculate memory consumption in step 3 as well as some custom consumption functions for certain {role, segment} pairs.

### Step 2 Generic Rule Excluding Static Service Pools Wizard

For every {role, segment} pair where the segment defines a default value, the pair's minimum is set to the segment's minimum value (or 0 if undefined), and the ideal is set to the segment's default value.

### Step 2 Custom Rules Excluding Static Service Pools Wizard

## HDFS

For the NameNode and Secondary NameNode JVM heaps, the minimum is 50 MB and the ideal is  $\max(1 \text{ GB}, \text{sum\_over\_all}(\text{DataNode mountpoints' available space}) / 0.000008)$ .

## Configuring CDH and Managed Services

### MapReduce

For the JobTracker JVM heap, the minimum is 50 MB and the ideal is  $\max(1 \text{ GB}, \text{round}((1 \text{ GB} * 2.3717181092 * \ln(\text{number of TaskTrackers in MapReduce service})) - 2.6019933306))$ . If there are  $\leq 5$  TaskTrackers, the ideal is 1 GB.

For the mapper JVM heaps, the minimum is 1 and the ideal is (number of cores, including hyperthreads, on the TaskTracker host). Note that memory consumption is scaled by `mapred_child_java_opts_max_heap` (the size of a given task's heap).

For the reducer JVM heaps, the minimum is 1 and the ideal is (number of cores, including hyperthreads, on the TaskTracker host) / 2. Note that memory consumption is scaled by `mapred_child_java_opts_max_heap` (the size of a given task's heap).

### YARN

For the memory total allowed for containers, the minimum is 1 GB and the ideal is  $\min(8 \text{ GB}, (\text{total RAM on NodeManager host}) * 0.8)$ .

### Hue

With the exception of the Beeswax Server (present only in CDH 4), Hue roles don't have memory limits. Therefore, Cloudera Manager treats them as roles that consume a fixed amount of memory by setting their minimum and ideal consumption values, but not their configuration values. The two consumption values are set to 256 MB.

### Impala

With the exception of the Impala Daemon, Impala roles don't have memory limits. Therefore Cloudera Manager treats them as roles that consume a fixed amount of memory by setting their minimum/ideal consumption values, but not their configuration values. The two consumption values are set to 150 MB for the Catalog Server and 64 MB for the StateStore.

For the Impala Daemon memory limit, the minimum is 256 MB and the ideal is  $((\text{total RAM on daemon host}) * 0.64)$ .

### Solr

For the Solr Server JVM heap, the minimum is 50 MB and the ideal is  $(\min(64 \text{ GB}, (\text{total RAM on Solr Server host}) * 0.64) / 2.6)$ . For the Solr Server JVM direct memory segment, the minimum is 256 MB and the ideal is  $(\min(64 \text{ GB}, (\text{total RAM on Solr Server host}) * 0.64) / 2)$ .

### Cloudera Management Service

- Alert Publisher JVM heap - treated as if it consumed a fixed amount of memory by setting the minimum/ideal consumption values, but not the configuration values. The two consumption values are set to 256 MB.
- Service and Host Monitor JVM heaps - the minimum is 50 MB and the ideal is either 256 MB (10 or fewer managed hosts), 1 GB (100 or fewer managed hosts), or 2 GB (over 100 managed hosts).
- Event Server, Reports Manager, and Navigator Audit Server JVM heaps - the minimum is 50 MB and the ideal is 1 GB.
- Navigator Metadata Server JVM heap - the minimum is 512 MB and the ideal is 2 GB.
- Service and Host Monitor off-heap memory segments - the minimum is either 768 MB (10 or fewer managed hosts), 2 GB (100 or fewer managed hosts), or 6 GB (over 100 managed hosts). The ideal is always twice the minimum.

#### Step 2 Generic Rule for Static Service Pools Wizard

For every {role, segment} pair where the segment defines a default value and an autoconfiguration share, the pair's minimum is set to the segment's default value, and the ideal is set to  $\min((\text{segment soft max (if exists) or segment max (if exists) or } 2^{63}-1), (\text{total RAM on role's host} * 0.8 / \text{segment scale factor} * \text{service percentage chosen in wizard} * \text{segment autoconfiguration share}))$ .

Autoconfiguration shares are defined as follows:

- HBase RegionServer JVM heap: 1
- HDFS DataNode JVM heap: 1 in CDH 4, 0.2 in CDH 5
- HDFS DataNode maximum locked memory: 0.8 (CDH 5 only)
- Solr Server JVM heap: 0.5
- Solr Server JVM direct memory: 0.5
- Spark Worker JVM heap: 1
- Accumulo Tablet Server JVM heap: 1
- Add-on services: any

Roles not mentioned here do not define autoconfiguration shares and thus aren't affected by this rule.

Additionally, there's a generic rule to handle `cgroup.memory_limit_in_bytes`, which is unused by Cloudera services but is available for add-on services. Its behavior varies depending on whether the role in question has segments or not.

### With Segments

The minimum is the `min(cgroup.memory_limit_in_bytes_min (if exists) or 0, sum_over_all(segment minimum consumption))`, and the ideal is the sum of all segment ideal consumptions.

### Without Segments

The minimum is `cgroup.memory_limit_in_bytes_min (if exists) or 0`, and the ideal is `(total RAM on role's host * 0.8 * service percentage chosen in wizard)`.

## Step 3 Custom Rules for Static Service Pools Wizard

### YARN

For the memory total allowed for containers, the minimum is 1 GB and the ideal is `min(8 GB, (total RAM on NodeManager host) * 0.8 * service percentage chosen in wizard)`.

### Impala

For the Impala Daemon memory limit, the minimum is 256 MB and the ideal is `((total RAM on Daemon host) * 0.8 * service percentage chosen in wizard)`.

### MapReduce

- mapper JVM heaps - the minimum is 1 and the ideal is `(number of cores, including hyperthreads, on the TaskTracker host * service percentage chosen in wizard)`. Note that memory consumption is scaled by `mapred_child_java_opts_max_heap` (the size of a given task's heap).
- reducer JVM heaps - the minimum is 1 and the ideal is `(number of cores, including hyperthreads on the TaskTracker host * service percentage chosen in wizard) / 2`. Note that memory consumption is scaled by `mapred_child_java_opts_max_heap` (the size of a given task's heap).

## Step 3 Generic Rule

For every {role, segment} pair, the segment's current value is converted into bytes, and then multiplied by the scale factor (1.0 by default, 1.3 for JVM heaps, and freely defined for Custom Service Descriptor services).

## Step 3 Custom Rules

## Configuring CDH and Managed Services

### Impala

For the Impala Daemon, the memory consumption is 0 if `YARN_For_ResourceManager` is set. If the memory limit is defined but not -1, its value is used verbatim. If it's defined but -1, the consumption is equal to the total RAM on the Daemon host. If it is undefined, the consumption is (total RAM \* 0.8).

### MapReduce

See [Step 3 Custom Rules for Static Service Pools Wizard](#) on page 15.

### Solr

For the Solr Server JVM direct memory segment, the consumption is equal to the value verbatim provided `solr.hdfs.blockcache.enable` and `solr.hdfs.blockcache.direct.memory.allocation` are both true. Otherwise, the consumption is 0.

## Step 7 Custom Rules

### HDFS

- NameNode JVM heaps are equalized. For every pair of NameNodes in an HDFS service with different heap sizes, the larger heap size is reset to the smaller one.
- JournalNode JVM heaps are equalized. For every pair of JournalNodes in an HDFS service with different heap sizes, the larger heap size is reset to the smaller one.
- NameNode and Secondary NameNode JVM heaps are equalized. For every {NameNode, Secondary NameNode} pair in an HDFS service with different heap sizes, the larger heap size is reset to the smaller one.

### HBase

Master JVM heaps are equalized. For every pair of Masters in an HBase service with different heap sizes, the larger heap size is reset to the smaller one.

### MapReduce

JobTracker JVM heaps are equalized. For every pair of JobTrackers in an MapReduce service with different heap sizes, the larger heap size is reset to the smaller one.

### Oozie

Oozie Server JVM heaps are equalized. For every pair of Oozie Servers in an Oozie service with different heap sizes, the larger heap size is reset to the smaller one.

### YARN

ResourceManager JVM heaps are equalized. For every pair of ResourceManagers in a YARN service with different heap sizes, the larger heap size is reset to the smaller one.

### ZooKeeper

ZooKeeper Server JVM heaps are equalized. For every pair of servers in a ZooKeeper service with different heap sizes, the larger heap size is reset to the smaller one.

### Impala

If an Impala service has `YARN_For_ResourceManager` set, every Impala Daemon memory limit is set to the value of (`yarn.nodemanager.resource.memory-mb` \* 1 GB) if there's a YARN NodeManager co-located with the Impala Daemon.



## General Rules

### HDFS

- `dfs.datanode.du.reserved` - For each DataNode, set to  $\min((\text{total space of DataNode host largest mountpoint}) / 10, 10 \text{ GB})$ .
- `dfs.namenode.name.dir` - For each NameNode, set to the first two mountpoints on the NameNode host with `/dfs/nn` appended.
- `dfs.namenode.checkpoint.dir` - For each Secondary NameNode, set to the first mountpoint on the Secondary NameNode host with `/dfs/snn` appended.
- `dfs.datanode.data.dir` - For each DataNode, set to all the mountpoints on the host with `/dfs/dn` appended.
- `dfs.journalnode.edits.dir` - For each JournalNode, set to the first mountpoint on the JournalNode host with `/dfs/jn` appended.
- `dfs.datanode.failed.volumes.tolerated` - For each DataNode, set to  $(\text{number of mountpoints on DataNode host}) / 2$ .
- `dfs.namenode.service.handler.count` and `dfs.namenode.handler.count` - For each NameNode, set to  $\max(30, \ln(\text{number of DataNodes in this HDFS service}) * 20)$ .
- `dfs.block.local-path-access.user` - For each HDFS service, set to `impala` if there's an Impala service in the cluster. *This rule is unscoped; it can fire even if the HDFS service is not under scope.*
- `dfs.datanode.hdfs-blocks-metadata.enabled` - For each HDFS service, set to `true` if there's an Impala service in the cluster. *This rule is unscoped; it can fire even if the HDFS service is not under scope.*
- `dfs.client.read.shortcircuit` - For each HDFS service, set to `true` if there's an Impala service in the cluster. *This rule is unscoped; it can fire even if the HDFS service is not under scope.*
- `dfs.datanode.data.dir.perm` - For each DataNode, set to `755` if there's an Impala service in the cluster and the cluster isn't Kerberized. *This rule is unscoped; it can fire even if the HDFS service is not under scope.*
- `fs.trash.interval` - For each HDFS service, set to `1`.

### MapReduce

- `mapred.local.dir` - For each JobTracker, set to the first mountpoint on the JobTracker host with `/mapred/jt` appended.
- `mapred.local.dir` - For each TaskTracker, set to all the mountpoints on the host with `/mapred/local` appended.
- `mapred.reduce.tasks` - For each MapReduce service, set to  $\max(1, \text{sum\_over\_all}(\text{TaskTracker number of reduce tasks (determined via mapred.tasktracker.reduce.tasks.maximum for that TaskTracker, which is configured separately)}) / 2)$ .
- `mapred.job.tracker.handler.count` - For each JobTracker, set to  $\max(10, \ln(\text{number of TaskTrackers in this MapReduce service}) * 20)$ .
- `mapred.submit.replication` - If there's an HDFS service in the cluster, for each MapReduce service, set to  $\max(1, \sqrt{\text{number of DataNodes in the HDFS service}})$ .
- `mapred.tasktracker.instrumentation` - If there's a management service, for each MapReduce service, set to `org.apache.hadoop.mapred.TaskTrackerCmonInst`. *This rule is unscoped; it can fire even if the MapReduce service is not under scope.*

### HBase

- `hbase.replication` - For each HBase service, set to `true` if there's a Key-Value Store Indexer service in the cluster. *This rule is unscoped; it can fire even if the HBase service is not under scope.*
- `replication.replicationsource.implementation` - For each HBase service, set to `com.ngdata.sep.impl.SepReplicationSource` if there's a Keystore Indexer service in the cluster. *This rule is unscoped; it can fire even if the HBase service is not under scope.*

## Configuring CDH and Managed Services

### YARN

- `yarn.nodemanager.local-dirs` - For each NodeManager, set to all the mountpoints on the NodeManager host with `/yarn/nm` appended.
- `yarn.nodemanager.resource.cpu-vcores` - For each NodeManager, set to the number of cores (including hyperthreads) on the NodeManager host.
- `mapred.reduce.tasks` - For each YARN service, set to  $\max(1, \text{sum\_over\_all}(\text{NodeManager number of cores, including hyperthreads}) / 2)$ .
- `yarn.resourcemanager.nodemangers.heartbeat-interval-ms` - For each NodeManager, set to  $\max(100, 10 * (\text{number of NodeManagers in this YARN service}))$ .
- `yarn.scheduler.maximum-allocation-vcores` - For each ResourceManager, set to  $\max\_over\_all(\text{NodeManager number of vcores (determined via } \text{yarn.nodemanager.resource.cpu-vcores} \text{ for that NodeManager, which is configured separately)})$ .
- `yarn.scheduler.maximum-allocation-mb` - For each ResourceManager, set to  $\max\_over\_all(\text{NodeManager amount of RAM (determined via } \text{yarn.nodemanager.resource.memory-mb} \text{ for that NodeManager, which is configured separately)})$ .
- `mapreduce.client.submit.file.replication` - If there's an HDFS service in the cluster, for each YARN service, set to  $\max(1, \text{sqrt}(\text{number of DataNodes in the HDFS service}))$ .

### Hue

- **WebHDFS dependency** - For each Hue service, set to either the first HttpFS role in the cluster, or, if there are none, the first NameNode in the cluster.
- **HBase Thrift Server dependency** - For each Hue service in a CDH 4.4 or later cluster, set to the first HBase Thrift Server in the cluster.

### Impala

For each Impala service, set **Enable Impala auditing for Navigator** to true if there's a Cloudera Management Service with a Navigator audit server role. *This rule is unscoped; it can fire even if the Impala service is not under scope.*

### All Services

If a service dependency is unset, and a service with the desired type exists in the cluster, set the service dependency to the first such target service. Applies to all service dependencies except `YARN_For_ResourceManager`. Applies only to the Express and Add Cluster wizards.

### Role-Host Placement

Cloudera Manager employs the same role-host placement rule regardless of wizard. The set of hosts considered depends on the scope. If the scope is a cluster, all hosts in the cluster are included. If a service, all hosts in the service's cluster are included. If the Cloudera Management Service, all hosts in the deployment are included. The rules are as follows:

1. The hosts are sorted from most to least physical RAM. Ties are broken by sorting on hostname (ascending) followed by host identifier (ascending).
2. The overall number of hosts is used to determine which arrangement to use. These arrangements are hard-coded, each dictating for a given "master" role type, what index (or indexes) into the sorted host list in step 1 to use.
3. Master role types are included based on several factors:
  - Is this role type part of the service (or services) under scope?
  - Does the service already have the right number of instances of this role type?
  - Does the cluster's CDH version support this role type?
  - Does the installed Cloudera Manager license allow for this role type to exist?

4. Master roles are placed on each host using the indexes and the sorted host list. If a host already has a given master role, it is skipped.
5. An HDFS DataNode is placed on every host outside of the arrangement described in step 2, provided HDFS is one of the services under scope.
6. Certain "worker" roles are placed on every host where an HDFS DataNode exists, either because it existed there prior to the wizard, or because it was added in the previous step. The supported worker role types are:
  - MapReduce TaskTrackers
  - YARN NodeManagers
  - HBase RegionServers
  - Impala Daemons
  - Spark Workers
7. Hive gateways are placed on every host, provided a Hive service is under scope and a gateway didn't already exist on a given host.

This rule merely dictates the *default* placement of roles; you are free to modify it before it is applied by the wizard.

## Custom Configuration

**Required Role:** Configurator Cluster Administrator Full Administrator

Cloudera Manager exposes properties that allow you to insert custom configuration text into XML configuration, property, and text files, or into an environment. The naming convention for these properties is: **XXX Advanced Configuration Snippet (Safety Valve)** for **YYY** or **XXX YYY Advanced Configuration Snippet (Safety Valve)**, where **XXX** is a service or role and **YYY** is the target.

The values you enter into a configuration snippet must conform to the syntax of the target. For an XML configuration file, the configuration snippet must contain valid XML property definitions. For a properties file, the configuration snippet must contain valid property definitions. Some files simply require a list of host addresses.

The configuration snippet mechanism is intended for use in cases where there is configuration setting that is not exposed as a configuration property in Cloudera Manager. Configuration snippets generally override normal configuration. Contact Cloudera Support if you are required to use a configuration snippet that is not explicitly documented.

Service-wide configuration snippets apply to all roles in the service; a configuration snippet for a role group applies to all instances of the role associated with that role group.

There are configuration snippets for servers and client configurations. In general after changing a server configuration snippet you must [restart](#) the server, and after changing a client configuration snippet you must [redploy the client configuration](#). Sometimes you can refresh instead of restart. In some cases however, you must restart a dependent server after changing a client configuration. For example, changing a MapReduce client configuration marks the dependent Hive server as [stale](#), which must be restarted. The Admin Console displays an indicator when a server must be restarted. In addition, the All Configuration Issues tab on the [Home](#) page lists the actions you must perform to propagate the changes.

## Configuration Snippet Types and Syntax

Type	Description	Syntax
Configuration	Set configuration properties in various configuration files; the property name indicates into which configuration file the configuration will be placed. Configuration files have the extension <code>.xml</code> or <code>.conf</code> .	<pre>&lt;property&gt;   &lt;name&gt;property_name&lt;/name&gt;   &lt;value&gt;property_value&lt;/value&gt; &lt;/property&gt;</pre>

Type	Description	Syntax
	<p>For example, there are several configuration snippets for the Hive service. One Hive configuration snippet property is called the <b>HiveServer2 Advanced Configuration Snippet for hive-site.xml</b>; configuration you enter here is inserted verbatim into the <code>hive-site.xml</code> file associated with the HiveServer2 role group.</p> <p>To see a list of configuration snippets that apply to a specific configuration file, enter the configuration file name in the Search field in the top navigation bar. For example, searching for <code>mapred-site.xml</code> shows the configuration snippets that have <code>mapred-site.xml</code> in their name.</p>	<p>For example, to specify a MySQL connector library, put this property definition in that configuration snippet:</p> <pre>&lt;property&gt;   &lt;name&gt;hive.aux.jars.path&lt;/name&gt;   &lt;value&gt;file:///usr/share/java/mysql-connector-java.jar&lt;/value&gt; &lt;/property&gt;</pre>
Environment	<p>Specify key-value pairs for a service, role, or client that are inserted into the respective environment.</p> <p>One example of using an environment configuration snippet is to add a JAR to a classpath. Place JARs in a custom location such as <code>/opt/myjars</code> and extend the classpath via the appropriate service environment configuration snippet. The value of a JAR property must conform to the syntax supported by its environment. See <a href="#">Setting the class path</a>.</p> <p>Do not place JARs inside locations such as <code>/opt/cloudera</code> or <code>/usr/lib/{hadoop*,hbase*,hive*,etc.}</code> that are managed by Cloudera because they are overwritten at upgrades.</p>	<p><code>key=value</code></p> <p>For example, to add JDBC connectors to a Hive gateway classpath, add</p> <pre>AUX_CLASSPATH=/usr/share/java/mysql-connector-java.jar:/usr/share/java/odbc-connector-java.jar</pre> <p>or</p> <pre>AUX_CLASSPATH=/usr/share/java/*</pre> <p><b>to Gateway Client Advanced Configuration Snippet for hive-env.sh.</b></p>
Logging	Set properties in a <code>log4j.properties</code> file.	<pre>key1=value1 key2=value2</pre> <p>For example:</p> <pre>max.log.file.size=200MB max.log.file.backup.index=10</pre>
Metrics	Set properties to configure Hadoop metrics in a <code>hadoop-metrics.properties</code> or <code>hadoop-metrics2.properties</code> file.	<pre>key1=value1 key2=value2</pre> <p>For example:</p> <pre>*.sink.foo.class=org.apache.hadoop.metrics2.sink.FileSink namenode.sink.foo.filename=/tmp/namenode-metrics.out secondarynamenode.sink.foo.filename=/tmp/secondarynamenode-metrics.out</pre>
White and black lists	Specify a list of host addresses that are allowed or disallowed from accessing a service.	<pre>host1.domain1 host2.domain2</pre>

## Setting an Advanced Configuration Snippet

1. Click a service.
2. Click the **Configuration** tab.
3. Expand a category group and click an **Advanced** subcategory.

4. In the Property column, choose a property that contains the string **Advanced Configuration Snippet (Safety Valve)**.
5. Click the **Value** column to enable editing.
6. Specify the properties.
7. Click **Save Changes**.
8. Restart the service or role or redeploy client configurations as indicated.

## Stale Configurations

**Required Role:** Configurator Cluster Administrator Full Administrator




The Stale Configurations page provides differential views of changes made in a cluster. For any configuration change, the page contains entries of all affected attributes. For example, the following File entry shows the change to the file `hdfs-site.xml` when you update the property controlling how much disk space is reserved for non-HDFS use on each DataNode:

File: hdfs-site.xml		hdfs (3) Show
...	... @@ -91,9 +91,9 @@	
91	91 <value>4096</value>	
92	92 </property>	
93	93 <property>	
94	94 <name>dfs.datanode.du.reserved</name>	
95	95 - <value>5077964390</value>	
95	95 + <value>2147483648</value>	
96	96 </property>	
97	97 <property>	
98	98 <name>dfs.datanode.failed.volumes.tolerated</name>	
99	99 <value>0</value>	

To display the entities affected by a change, click the **Show** button at the right of the entry. The following dialog shows that three DataNodes were affected by the disk space change:

Entities Affected By This Change		×
Changes From: File: hdfs-site.xml		
<input type="text" value="Search Roles"/>		
<b>hdfs</b> (3) <ul style="list-style-type: none"> <li>datanode (tcdn48-4)</li> <li>datanode (tcdn48-2)</li> <li>datanode (tcdn48-3)</li> </ul>		
		Close

## Viewing Stale Configurations

To view stale configurations, click the , , or  indicator next to a service on the [Home Page](#) or on a service status page.

## Attribute Categories

The categories of attributes include:

- **Environment** - represents environment variables set for the role. For example, the following entry shows the change to the environment that occurs when you update the heap memory configuration of the SecondaryNameNode.

Environment		hdfs (1) Show
...	... @@ -2,6 +2,6 @@	
2	2 HADOOP_AUDIT_LOGGER=INFO,RFAUDIT	
3	3 HADOOP_LOGFILE=hadoop-cmf-HDFS-1-SECONDARYNAMENODE-tcdn48-1.ent.cloudera.com.log.out	
4	4 HADOOP_LOG_DIR=/var/log/hadoop-hdfs	
5	5 HADOOP_ROOT_LOGGER=INFO,RFA	
6	6 -HADOOP_SECONDARYNAMENODE_OPTS=-Xms305135616 -Xmx305135616 -XX:+UseParNewGC -XX:+UseConcMarkSweepGC -XX:-CMSConcurrentMTEnabled -XX:CMSInitiatingOccupancy...	
6	6 +HADOOP_SECONDARYNAMENODE_OPTS=-Xms1073741824 -Xmx1073741824 -XX:+UseParNewGC -XX:+UseConcMarkSweepGC -XX:-CMSConcurrentMTEnabled -XX:CMSInitiatingOccupancy...	
7	7 HADOOP_SECURITY_LOGGER=INFO,RFA	

## Configuring CDH and Managed Services

- **Files** - represents configuration files used by the role.
- **Process User & Group** - represents the user and group for the role. Every role type has a configuration to specify the user/group for the process. If you change a value for a user or group on any service's configuration page it will appear in the Stale Configurations page.
- **System Resources** - represents system resources allocated for the role, including ports, directories, and cgroup limits. For example, a change to the port of role instance will appear in the System Resources category.
- **Client Configs Metadata** - represents client configurations.

### Filtering Stale Configurations


You filter the entries on the Stale Configurations page by selecting from one of the drop-down lists:

- **Attribute** - you can filter by an attribute category such as All Files or by a specific file such as `topology.map` or `yarn-site.xml`.
- **Service**
- **Role**

After you make a selection, both the page and the drop-down show only entries that match that selection.

To reset the view, click **Remove Filter** or select **All XXX**, where XXX is Files, Services, or Roles, from the drop-down. For example, to see all the files, select **All Files**.

### Actions

The Stale Configurations page displays action links. The action depends on what is required to bring the entire cluster's configuration up to date. If you navigate to the page by clicking a  (Refresh Needed) indicator, the action button will say **Restart Cluster** if *one* of the roles listed on the page need to be restarted.

- **Refresh Cluster** - Runs the [cluster refresh](#) action.
- **Restart Cluster** - Runs the [cluster restart](#) action.
- **Restart Cloudera Management Service** - Runs the [restart Cloudera Management Service](#) action.
- **Deploy Client Configuration** - Runs the [cluster deploy client configurations](#) action.

### Client Configuration Files


**Required Role:** Configurator Cluster Administrator Full Administrator

To allow clients to use the HBase, HDFS, Hive, MapReduce, and YARN services, Cloudera Manager creates zip archives of the configuration files containing the service properties. The zip archive is referred to as a **client configuration file**. Each archive contains the set of configuration files needed to access the service: for example, the MapReduce client configuration file contains copies of `core-site.xml`, `hadoop-env.sh`, `hdfs-site.xml`, `log4j.properties` and `mapred-site.xml`.

Client configuration files are generated automatically by Cloudera Manager based on the services and roles you have installed and Cloudera Manager deploys these configurations automatically when you install your cluster, add a service on a host, or add a [gateway role](#) on a host. Specifically, for each host that has a service role instance installed, and for each host that is configured as a gateway role for that service, the deploy function downloads the configuration zip file, unzips it into the appropriate configuration directory, and uses the Linux [alternatives](#) mechanism to set a given, configurable priority level. If you are installing on a system that happens to have pre-existing alternatives, then it is possible another alternative may have higher priority and will continue to be used. The alternatives priority of the Cloudera Manager client configuration is configurable under the **Gateway** sections of the **Configuration** tab for the appropriate service.

You can also manually distribute client configuration files to the clients of a service.

The main circumstance that may require a redeployment of the client configuration files is when you have modified a configuration. In this case you will typically see a message instructing you to redeploy your client

configurations. The affected service(s) will also display a  icon. Click the indicator to display the [Stale Configurations](#) on page 21 page.

### How Client Configurations are Deployed

Client configuration files are deployed on any host that is a client for a service—that is, that has a role for the service on that host. This includes roles such as DataNodes, TaskTrackers, RegionServers and so on as well as gateway roles for the service.


If roles for multiple services are running on the same host (for example, a DataNode role and a TaskTracker role on the same host) then the client configurations for both roles are deployed on that host, with the alternatives priority determining which configuration takes precedence.

For example, suppose we have six hosts running roles as follows: host H1: HDFS-NameNode; host H2: MR-JobTracker; host H3: HBase-Master; host H4: MR-TaskTracker, HDFS-DataNode, HBase-RegionServer; host H5: MR-Gateway; host H6: HBase-Gateway. Client configuration files will be deployed on these hosts as follows: host H1: hdfs-clientconfig (only); host H2: mapreduce-clientconfig; host H3: hbase-clientconfig; host H4: hdfs-clientconfig, mapreduce-clientconfig, hbase-clientconfig; host H5: mapreduce-clientconfig; host H6: hbase-clientconfig

If the HDFS NameNode and MapReduce JobTracker were on the same host, then that host would have both hdfs-clientconfig and mapreduce-clientconfig installed.

### Downloading Client Configuration Files

1. Follow the appropriate procedure according to your starting point:


Page	Procedure
Home	<ol style="list-style-type: none"> <li>1. On the Home page, click  to the right of the cluster name and select <b>View Client Configuration URLs</b>. A pop-up window with links to the configuration files for the services you have installed displays.</li> <li>2. Click a link or save the link URL and download the file using <code>wget</code> or <code>curl</code>.</li> </ol>
Service	<ol style="list-style-type: none"> <li>1. Go to a service whose client configuration you want to download.</li> <li>2. Select <b>Actions</b> &gt; <b>Download Client Configuration</b>.</li> </ol>

### Manually Redeploying Client Configuration Files

Although Cloudera Manager will deploy client configuration files automatically in many cases, if you have modified the configurations for a service, you may need to redeploy those configuration files.

If your client configurations were deployed automatically, the command described in this section will attempt to redeploy them as appropriate.

- **Note:** If you are deploying client configurations on a host that has multiple services installed, some of the same configuration files, though with different configurations, will be installed in the `conf` directories for each service. Cloudera Manager uses the `priority` parameter in the `alternatives --install` command to ensure that the correct configuration directory is made active based on the combination of services on that host. The priority order is YARN > MapReduce > HDFS. The priority can be configured under the **Gateway** sections of the **Configuration** tab for the appropriate service.

1. On the Home page, click  to the right of the cluster name and select **Deploy Client Configuration**.
2. Click **Deploy Client Configuration**.

### Viewing and Reverting Configuration Changes

Required Role: Configurator Cluster Administrator Full Administrator



- **Important: This feature is available only with a Cloudera Enterprise license.**

For other licenses, the following applies:

- Cloudera Express- The feature is not available.
- Cloudera Enterprise Data Hub Edition Trial - The feature is available until you end the trial or the trial license expires.

To obtain a license for Cloudera Enterprise, fill in this [form](#) or call 866-843-7207. After you install a Cloudera Enterprise license, the feature will be available.

Whenever you change and save a set of configuration settings for a service or role instance, or a host, Cloudera Manager saves a revision of the previous settings and the name of the user who made the changes. You can then view past revisions of the configuration settings, and, if desired, roll back the settings to a previous state.

### Viewing Configuration Changes

1. For a service, role, or host, and click the **Configuration** tab and the **History and Rollback** button. The most recent revision, currently in effect, is shown under **Current Revision**. Prior revisions are shown under **Past Revisions**.
  - By default, or if you click **Show All**, a list of all revisions is shown. If you are viewing a service or role instance, all service/role group related revisions are shown. If you are viewing a host or all hosts, all host/all hosts related revisions are shown.
  - To list only the configuration revisions that were done in a particular time period, use the Time Range Selector to [select a time range](#). Then, click **Show within the Selected Time Range**.
2. Click the **Details...** link. The Revision Details dialog displays.

### Revision Details Dialog

For a service or role instance, shows the following:

- A brief message describing the context of the changes.
- The date/time stamp of the change.
- The user who performed the change.
- The names of any role groups created.
- The names of any role groups deleted.

For a host instance, shows just a message, date and time stamp, and the user.

The dialog contains two tabs:

- **Configuration Values** - displays configuration value changes, where changes are organized under the role group to which they were applied. (For example, if you changed a Service-Wide property, it will affect all role groups for that service). For each modified property, the Value column shows the new value of the property and the previous value.
- **Group Membership** - displays changes to the changed the group membership of a role instance (moved the instance from one group to another). This tab is only shown for service and role configurations.

### Reverting Configuration Changes

1. Select the current or past revision to which to roll back.
2. Click the **Details...** link. The Revision Details dialog displays.
3. Click the **Configuration Values** tab.
4. Click the **Revert Configuration Changes** button. The revert action occurs immediately. You may need to restart the service or the affected roles for the change to take effect.




- **Important:** This feature can only be used to revert changes to configuration values. You cannot use this feature to:
  - Revert NameNode High Availability. You must perform this action by explicitly [disabling High Availability](#).
  - Disable [Kerberos security](#).
  - Revert role group actions (creating, deleting, or moving membership among groups). You must perform these actions explicitly in the [Role Groups](#) on page 42 feature.

## Exporting and Importing Cloudera Manager Configuration

You can use the Cloudera Manager API to programmatically export and import a definition of all the entities in your Cloudera Manager-managed deployment—clusters, service, roles, hosts, users and so on. See the [Cloudera Manager API](#) documentation on how to manage deployments using the [/cm/deployment](#) resource.

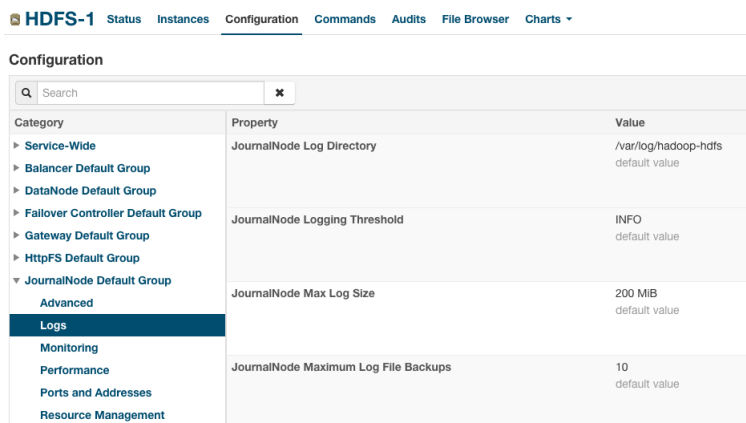
## Managing Clusters

Cloudera Manager can manage multiple clusters. Once you have successfully installed your first cluster, you can add additional clusters, running the same or a different version of CDH. You can then manage each cluster and its services independently.

On the Home page you can access many cluster-wide actions by selecting  to the right of the cluster name: add a service, start, stop, restart, deploy client configurations, enable Kerberos, and perform cluster refresh, rename, upgrade, and maintenance mode actions.

The Cloudera Manager configuration screens offer two layout options: classic and new. The classic layout is the default; however, on each configuration page you can easily switch between the classic and new layouts using the **Switch to XXX layout** link at the top right of the page. All the configuration procedures described in the Cloudera Manager documentation assume the classic layout.

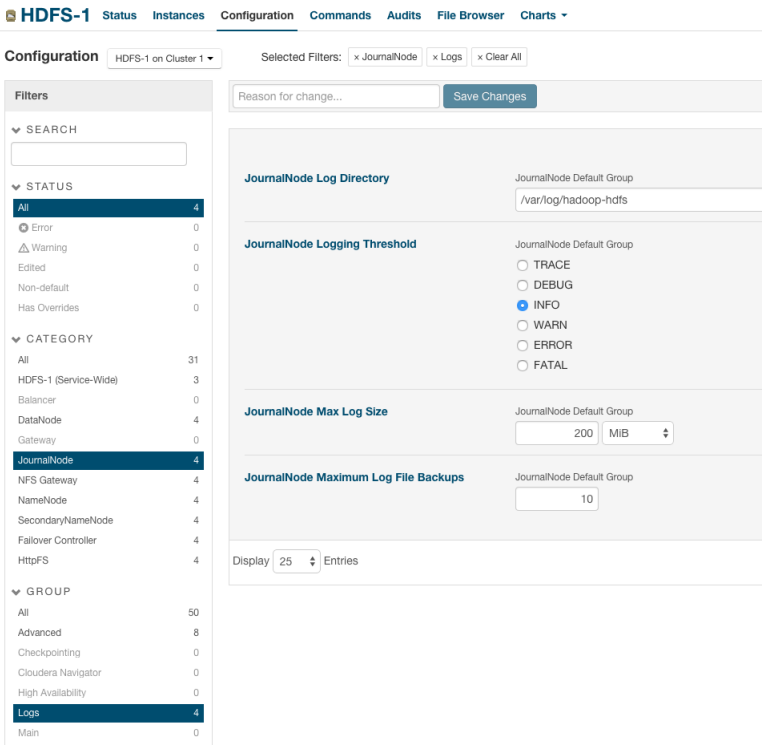
- **classic** - pages are organized by role group and categories within the role group. For example, to display the JournalNode maximum log size property (JournalNode Max Log Size), select **JournalNode Default Group > Logs**.



Category	Property	Value
Service-Wide	JournalNode Log Directory	/var/log/hadoop-hdfs default value
Balancer Default Group		
DataNode Default Group		
Failover Controller Default Group	JournalNode Logging Threshold	INFO default value
Gateway Default Group		
HttpFS Default Group		
JournalNode Default Group		
Advanced	JournalNode Max Log Size	200 MiB default value
Logs		
Monitoring		
Performance	JournalNode Maximum Log File Backups	10 default value
Ports and Addresses		
Resource Management		

When a configuration property has been set to a value different from the default, a **Reset to the default value** link displays.

- **new** - pages contain controls that allow you filter configuration properties based on configuration status, category, and group. For example, to display the JournalNode maximum log size property (JournalNode Max Log Size), click the **CATEGORY > JournalNode** and **GROUP > Logs** filters:




When a configuration property has been set to a value different from the default, a reset to default value icon displays.


Adding and Deleting Clusters

Required Role: Full Administrator


Cloudera Manager can manage multiple clusters. Furthermore, the clusters do not need to run the same version of CDH; you can manage both CDH 4 and CDH 5 clusters with Cloudera Manager.

Adding a Cluster

Action	Procedure
New Hosts	<ol style="list-style-type: none"><li>On the <b>Home</b> page, click  and select <b>Add Cluster</b>. This begins the Installation Wizard, just as if you were installing a cluster for the first time. (See <a href="#">Cloudera Manager Deployment</a> for detailed instructions.)</li><li>To find new hosts, not currently managed by Cloudera Manager, where you want to install CDH, enter the host names or IP addresses, and click <b>Search</b>. Cloudera Manager lists the hosts you can use to configure a new cluster. Managed hosts that already have services installed will not be selectable.</li><li>Click <b>Continue</b> to install the new cluster. At this point the installation continues through the wizard the same as it did when you installed your first cluster. You will be asked to select the version of CDH to install, which services you want and so on, just as previously.</li><li>Restart the Reports Manager role.</li></ol>
Managed Hosts	You may have hosts that are already "managed" but are not part of a cluster. You can have managed hosts that are not part of a cluster when you have added hosts to Cloudera Manager either through the Add Host wizard, or by manually installing the Cloudera Manager agent onto hosts where you have not install any other services. This will also be the case if you remove all services from a host so that it no longer is part of a cluster.

Action	Procedure
	<ol style="list-style-type: none"> <li>1. On the <b>Home</b> page, click  and select <b>Add Cluster</b>. This begins the Installation Wizard, just as if you were installing a cluster for the first time. (See <a href="#">Cloudera Manager Deployment</a> for detailed instructions.)</li> <li>2. To see the list of the currently managed hosts, click the <b>Currently Managed Hosts</b> tab. This tab does not appear if you have no currently managed hosts that are not part of a cluster.</li> <li>3. To perform the installation, click <b>Continue</b>. Instead of searching for hosts, this will attempt to install onto any hosts managed by Cloudera Manager that are not already part of a cluster. It will proceed with the installation wizard as for a new cluster installation.</li> <li>4. Restart the Reports Manager role.</li> </ol>


### Deleting a Cluster

1. [Stop](#) the cluster.
2. On the **Home** page, click  to the right of the cluster name and select **Delete**.

### Starting, Stopping, Refreshing, and Restarting a Cluster

Required Role: **Operator** **Configurator** **Cluster Administrator** **Full Administrator**


#### Starting a Cluster

1. On the Home page, click  to the right of the cluster name and select **Start**.
2. Click **Start** that appears in the next screen to confirm. The **Command Details** window shows the progress of starting services.

When **All services successfully started** appears, the task is complete and you can close the **Command Details** window.

- **Note:** The cluster-level Start action starts only CDH and other product services (Impala, Cloudera Search). It does not start the Cloudera Management Service. You must [start the Cloudera Management Service](#) separately if it is not already running.

#### Stopping a Cluster

1. On the Home page, click  to the right of the cluster name and select **Stop**.
2. Click **Stop** in the confirmation screen. The **Command Details** window shows the progress of stopping services.

When **All services successfully stopped** appears, the task is complete and you can close the **Command Details** window.

- **Note:** The cluster-level Stop action does not stop the Cloudera Management Service. You must [stop the Cloudera Management Service](#) separately.

#### Refreshing a Cluster

Runs a cluster refresh action to bring the configuration up to date without restarting all services. For example, certain masters (for example NameNode and ResourceManager) have some configuration files (for example, `fair-scheduler.xml`, `mapred_hosts_allow.txt`, `topology.map`) that can be refreshed. If anything changes in those files then a refresh can be used to update them in the master. Here is a summary of the operations performed in a refresh action:

## Configuring CDH and Managed Services

✓ **Refresh Cluster**

Cluster 1

Finished

Mar 19, 2014 11:31:55 AM PDT

Mar 19, 2014 11:32:09 AM PDT

Successfully refreshed roles in the cluster.

**Command Progress**

Completed 4 of 4 steps.

✓ Run 1 steps in parallel  
Successfully refreshed datanode allow/exclude lists.  
[Details](#) ⌵


✓ Run 1 steps in parallel  
Successfully refreshed ResourceManager.  
[Details](#) ⌵

✓ Run 3 steps in parallel  
Successfully refreshed NodeManager.  
[Details](#) ⌵

✓ Run 3 steps in parallel  
Refreshed Impala Daemon's Pools configuration and ACLs successfully.  
[Details](#) ⌵

To refresh a cluster, in the Home page, click  to the right of the cluster name and select **Refresh Cluster**.


### Restarting a Cluster

1. On the Home page, click  to the right of the cluster name and select **Restart**.
2. Click **Restart** that appears in the next screen to confirm. The **Command Details** window shows the progress of stopping services.

When **All services successfully started** appears, the task is complete and you can close the **Command Details** window.

### Renaming a Cluster

**Required Role:** **Full Administrator**

1. On the Home page, click  to the right of the cluster name and select **Rename Cluster**.
2. Type the new cluster name and click **Rename Cluster**.

### Cluster-Wide Configuration

**Required Role:** **Configurator** **Cluster Administrator** **Full Administrator**

To view or set the values of a class of properties for an entire cluster:

1. On the Home page, click a cluster name.
2. Select **Configuration > Property**, where *Property* is:
  - All Non-default Values
  - All Log Directories
  - All Disk Space Thresholds
  - All Port Configurations

### Moving a Host Between Clusters

**Required Role:** **Full Administrator**

Moving a host between clusters can be accomplished by:

1. Decommissioning the host (see [Decommissioning Role Instances](#) on page 41).

2. Removing all roles from the host (except for the Cloudera Manager management roles). See [Deleting Role Instances](#) on page 41.
3. Deleting the host from the cluster (see [Deleting Hosts](#) on page 109), specifically the section on removing a host from a cluster but leaving it available to Cloudera Manager.
4. Adding the host to the new cluster (see [Adding a Host to the Cluster](#) on page 101).
5. Adding roles to the host (optionally using one of the host templates associated with the new cluster). See [Adding a Role Instance](#) on page 40 and [Host Templates](#) on page 103.

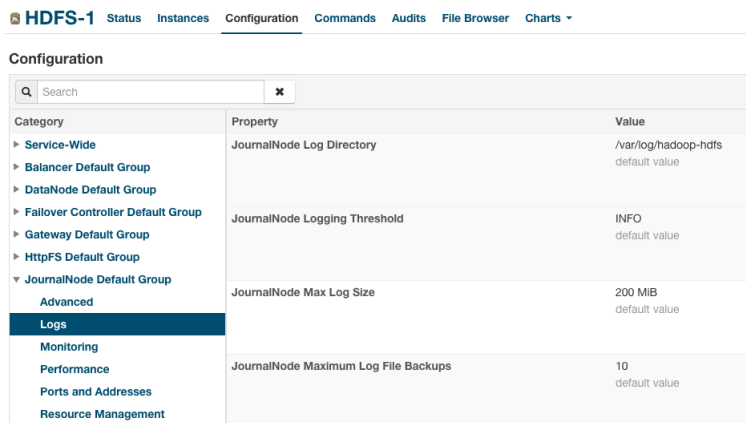
## Managing Services

Cloudera Manager service configuration features let you manage the deployment and configuration of CDH and managed services. You can add new services and roles if needed, gracefully start, stop and restart services or roles, and decommission and delete roles or services if necessary. Further, you can modify the configuration properties for services or for individual role instances. If you have a Cloudera Enterprise license, you can view past configuration changes and roll back to a previous revision. You can also generate client configuration files, enabling you to easily distribute them to the users of a service.

The topics in this chapter describe how to configure and use the services on your cluster. Some services have unique configuration requirements or provide unique features: those are covered in [Managing Individual Services](#) on page 43.

The Cloudera Manager configuration screens offer two layout options: classic and new. The classic layout is the default; however, on each configuration page you can easily switch between the classic and new layouts using the **Switch to XXX layout** link at the top right of the page. All the configuration procedures described in the Cloudera Manager documentation assume the classic layout.

- **classic** - pages are organized by role group and categories within the role group. For example, to display the JournalNode maximum log size property (JournalNode Max Log Size), select **JournalNode Default Group > Logs**.



Category	Property	Value
Service-Wide	JournalNode Log Directory	/var/log/hadoop-hdfs default value
Balancer Default Group		
DataNode Default Group		
Fallover Controller Default Group	JournalNode Logging Threshold	INFO default value
Gateway Default Group		
HttpFS Default Group		
JournalNode Default Group	JournalNode Max Log Size	200 MIB default value
Advanced		
Logs	JournalNode Maximum Log File Backups	10 default value
Monitoring		
Performance		
Ports and Addresses		
Resource Management		

When a configuration property has been set to a value different from the default, a **Reset to the default value** link displays.

- **new** - pages contain controls that allow you filter configuration properties based on configuration status, category, and group. For example, to display the JournalNode maximum log size property (JournalNode Max Log Size), click the **CATEGORY > JournalNode** and **GROUP > Logs** filters:

# Configuring CDH and Managed Services

**HDFS-1** Status Instances Configuration Commands Audits File Browser Charts ▾

**Configuration** HDFS-1 on Cluster 1 ▾ Selected Filters: x JournalNode x Logs x Clear All

**Filters**

▼ SEARCH

▼ STATUS

- All 4
- Error 0
- Warning 0
- Edited 0
- Non-default 0
- Has Overrides 0

▼ CATEGORY

- All 31
- HDFS-1 (Service-Wide) 3
- Balancer 0
- DataNode 4
- Gateway 0
- JournalNode 4**
- NFS Gateway 4
- NameNode 4
- SecondaryNameNode 4
- Fallover Controller 4
- HttpFS 4

▼ GROUP

- All 50
- Advanced 8
- Checkpointing 0
- Cloudera Navigator 0
- High Availability 0
- Logs 4**
- Main 0

Reason for change...

**Save Changes**

**JournalNode Log Directory** JournalNode Default Group  
/var/log/hadoop-hdfs

**JournalNode Logging Threshold** JournalNode Default Group

- ☐ TRACE
- ☐ DEBUG
- ☒ INFO
- ☐ WARN
- ☐ ERROR
- ☐ FATAL

**JournalNode Max Log Size** JournalNode Default Group  
200 MIB

**JournalNode Maximum Log File Backups** JournalNode Default Group  
10

Display 25 Entries

When a configuration property has been set to a value different from the default, a reset to default value icon displays.

## Adding a Service


**Required Role:** Full Administrator

After initial installation, you can use the **Add a Service** wizard to add and configure new service instances. For example, you may want to add a service such as Oozie that you did not select in the wizard during the initial installation.

You must use **Add a Service** to create a [Flume](#) service; because Flume requires you to specify the agent configuration, it must be added separately after the installation wizard has finished.

See [The MapReduce Service](#) on page 80 or [The YARN Service](#) on page 81 for information about how MapReduce and YARN services are related to each other.

To add a service:

1. On the Home page, click  to the right of the cluster name and select **Add a Service**. A list of service types display. You can add one type of service at a time.
2. Click the radio button next to the service to add and click **Continue**.
3. Select the radio button next to the services on which the new service should depend and click **Continue**.
4. Customize the assignment of role instances to hosts. The wizard evaluates the hardware configurations of the hosts to determine the best hosts for each role. The wizard assigns all worker roles to the same set of hosts to which the HDFS DataNode role is assigned. These assignments are typically acceptable, but you can reassign them if necessary.

Click a field below a role to display a dialog containing a list of hosts. If you click a field containing multiple hosts, you can also select **All Hosts** to assign the role to all hosts or **Custom** to display the pageable hosts dialog.

The following shortcuts for specifying hostname patterns are supported:

- Range of hostnames (without the domain portion)

Range Definition	Matching Hosts
10.1.1.[1-4]	10.1.1.1, 10.1.1.2, 10.1.1.3, 10.1.1.4
host[1-3].company.com	host1.company.com, host2.company.com, host3.company.com
host[07-10].company.com	host07.company.com, host08.company.com, host09.company.com, host10.company.com

- IP addresses
- Rack name

Click the **View By Host** button for an overview of the role assignment by hostname ranges.

5. Review and modify configuration settings, such as data directory paths and heap sizes and click **Continue**. The service is started.
6. Click **Continue** then click **Finish**. You are returned to the [Home](#) page.
7. Verify the new service is started properly by checking the health status for the new service. If the Health Status is **Good**, then the service started properly.

## Add-on Services

**Required Role:** Full Administrator

Cloudera Manager supports adding new types of services (referred to as an **add-on service**) to Cloudera Manager, allowing such services to leverage Cloudera Manager distribution, configuration, monitoring, resource management, and life-cycle management features. An add-on service can be provided by Cloudera or an independent software vendor (ISV). If you have multiple clusters managed by Cloudera Manager, an add-on service can be deployed on any of the clusters.

- **Note:** If the add-on service is already installed and running on hosts that are not currently being managed by Cloudera Manager, you must first add the hosts to a cluster that's under management. See [Adding a Host to the Cluster](#) on page 101 for details.

## Custom Service Descriptor Files

Integrating an add-on service requires a Custom Service Descriptor (CSD) file. A CSD file contains all the configuration needed to describe and manage a new service. A CSD is provided in the form of a JAR file.

Depending on the service, the CSD and associated software may be provided by Cloudera or by an ISV. The integration process assumes that the add-on service software (parcel or package) has been installed and is present on the cluster. The recommended method is for the ISV to provide the software as a parcel, but the actual mechanism for installing the software is up to the ISV. The instructions in [Installing an Add-on Service](#) on page 32 assume that you have obtained the CSD file from the Cloudera repository or from an ISV. It also assumes you have obtained the service software, ideally as a parcel, and have or will install it on your cluster either prior to installing the CSD or as part of the CSD installation process.

## Configuring the Location of Custom Service Descriptor Files

The default location for CSD files is `/opt/cloudera/csd`. You can change the location in the Cloudera Manager Admin Console as follows:

1. Select **Administration > Settings**.
2. Click the **Custom Service Descriptors** category.
3. Edit the **Local Descriptor Repository Path** property.
4. Click **Save Changes** to commit the changes.

## Configuring CDH and Managed Services


### Installing an Add-on Service

An ISV may provide its software in the form of a parcel, or they may have a different way of installing their software onto your cluster. If their installation process is not via a parcel, then you should install their software before adding the CSD file. Follow the instructions from the ISV for installing the software, if you have not done so already. If the ISV has provided their software as a parcel, they may also have included the location of their parcel repository in the CSD they have provided. In that case, install the CSD first and then install the parcel.

### Installing the Custom Service Descriptor File




1. Acquire the CSD file from Cloudera or an ISV.
2. Log on to the Cloudera Manager Server host, and place the CSD file under the location configured for CSD files.
3. Restart the Cloudera Manager Server:

```
service cloudera-scm-server restart
```


4. Log into the Cloudera Manager Admin Console and restart the Cloudera Management Service.
  - a. Do one of the following:
    - 1. Select **Clusters** > **Cloudera Management Service** > **Cloudera Management Service**.
    - 2. Select **Actions** > **Restart**.
    - On the Home page, click  to the right of **Cloudera Management Service** and select **Restart**.
  - b. Click **Restart** to confirm. The **Command Details** window shows the progress of stopping and then starting the roles.
  - c. When **Command completed with n/n successful subcommands** appears, the task is complete. Click **Close**.

### Installing the Parcel

If you have already installed the external software onto your cluster, you can skip these steps and proceed to [Adding an Add-on Service](#) on page 33.

1. Click  in the main navigation bar. If the vendor has included the location of the repository in the CSD, the parcel should already be present and ready for downloading. If the parcel is available, skip to [step 7](#).
2. Use one of the following methods to open the parcel settings page:
  - **Navigation bar**
    1. Click  in the top navigation bar.
    2. Click the **Edit Settings** button.
  - **Menu**
    1. Select **Administration** > **Settings**.
    2. Click the **Parcels** category.
  - **Tab**
    1. Click the **Hosts** tab.
    2. Click the **Configuration** tab.
    3. Click the **Parcels** category.
    4. Click the **Edit Settings** button.
3. In the **Remote Parcel Repository URLs** list, click  to open an additional row.
4. Enter the path to the repository.
5. Click **Save Changes** to commit the changes.



6. Click . The external parcel should appear in the set of parcels available for download.
7. Download, distribute, and activate the parcel. See [Managing Parcels](#).

### Adding an Add-on Service

Add the service following the procedure in [Adding a Service](#) on page 30.

### Uninstalling an Add-on Service

1. Stop all instances of the service.
2. Delete the service from all clusters. If there are other services that depend on the service you are trying to delete, you must delete those services first.
3. Log on to the Cloudera Manager Server host and remove the CSD file.
4. Restart the Cloudera Manager Server:

```
service cloudera-scm-server restart
```

5. After the server has restarted, log into the Cloudera Manager Admin Console and restart the Cloudera Management Service.
6. Optionally remove the parcel.

## Starting, Stopping, and Restarting Services


Required Role: **Operator** **Configurator** **Cluster Administrator** **Full Administrator**

### Starting and Stopping Services

It's important to start and stop services that have dependencies in the correct order. For example, because MapReduce and YARN have a dependency on HDFS, you must start HDFS before starting MapReduce or YARN. The Cloudera Management Service and Hue are the only two services on which no other services depend; although you can start and stop them at anytime, their preferred order is shown in the following procedures.

The Cloudera Manager cluster actions start and stop services in the correct order. To start or stop all services in a cluster, follow the instructions in [Starting, Stopping, Refreshing, and Restarting a Cluster](#) on page 27.

### Starting a Service on All Hosts


1. On the Home page, click  to the right of the service name and select **Start**.
2. Click **Start** that appears in the next screen to confirm. When you see a **Finished** status, the service has started.

The order in which to start services is:

1. Cloudera Management Service
2. ZooKeeper
3. HDFS
4. Solr
5. Flume
6. HBase
7. Key-Value Store Indexer
8. MapReduce or YARN
9. Hive
10. Impala
11. Oozie
12. Sqoop
13. Hue

- **Note:** If you are unable to start the HDFS service, it's possible that one of the roles instances, such as a DataNode, was running on a host that is no longer connected to the Cloudera Manager Server host, perhaps because of a hardware or network failure. If this is the case, the Cloudera Manager Server will be unable to connect to the Cloudera Manager Agent on that disconnected host to start the role instance, which will prevent the HDFS service from starting. To work around this, you can stop all services, abort the pending command to start the role instance on the disconnected host, and then restart all services again without that role instance. For information about aborting a pending command, see [Aborting a Pending Command](#) on page 37.

### Stopping a Service on All Hosts


1. On the Home page, click  to the right of the service name and select **Stop**.
2. Click **Stop** that appears in the next screen to confirm. When you see a **Finished** status, the service has stopped.

The order in which to stop services is:

1. Hue
2. Sqoop
3. Oozie
4. Impala
5. Hive
6. MapReduce or YARN
7. Key-Value Store Indexer
8. HBase
9. Flume
10. Solr
11. HDFS
12. ZooKeeper
13. Cloudera Management Service

### Restarting a Service

It is sometimes necessary to restart a service, which is essentially a combination of stopping a service and then starting it again. For example, if you change the hostname or port where the Cloudera Manager is running, or you enable TLS security, you must restart the Cloudera Management Service to update the URL to the Server.

1. On the Home page, click  to the right of the service name and select **Restart**.
2. Click **Start** on the next screen to confirm. When you see a **Finished** status, the service has restarted.

To restart all services, use the [restart cluster](#) action.

### Rolling Restart

Required Role: **Operator** **Configurator** **Cluster Administrator** **Full Administrator**

- **Important:** This feature is available only with a Cloudera Enterprise license.

For other licenses, the following applies:

- Cloudera Express- The feature is not available.
- Cloudera Enterprise Data Hub Edition Trial - The feature is available until you end the trial or the trial license expires.

To obtain a license for Cloudera Enterprise, fill in this [form](#) or call 866-843-7207. After you install a Cloudera Enterprise license, the feature will be available.

Rolling restart allows you to conditionally restart the role instances of Flume, HBase, HDFS, MapReduce, YARN, and ZooKeeper services to update software or use a new configuration. If the service is not running, rolling restart is not available for that service. You can do a rolling restart of each service individually.

If you have [HDFS high availability](#) enabled, you can also perform a cluster-level rolling restart. At the cluster level, the rolling restart of worker hosts is performed on a host-by-host basis, rather than per service, to avoid all roles for a service potentially being unavailable at the same time. During a cluster restart, in order to avoid having your NameNode (and thus the cluster) being unavailable during the restart, Cloudera Manager will force a failover to the standby NameNode.

[MapReduce \(MRv1\) JobTracker High Availability](#) on page 245 and [YARN \(MRv2\) ResourceManager High Availability](#) on page 237 is *not* required for a cluster-level rolling restart. However, if you have JobTracker or ResourceManager high availability enabled, Cloudera Manager will force a failover to the standby JobTracker or ResourceManager.

### Performing a Service or Role Rolling Restart

You can initiate a rolling restart from either the Status page for one of the eligible services, or from the service's Instances page, where you can select individual roles to be restarted.

1. Go to the service you want to restart.
2. Do one of the following:
  - **service** - Select **Actions > Rolling Restart**.
  - **role** -
    1. Click the **Instances** tab.
    2. Select the roles to restart.
    3. Select **Actions for Selected > Rolling Restart**.
3. In the pop-up dialog box, select the options you want:
  - Restart only roles whose configurations are stale
  - Restart only roles that are running outdated software versions
  - Which role types to restart
4. If you select an HDFS, HBase, MapReduce, or YARN service, you can have their worker roles restarted in batches. You can configure:
  - How many roles should be included in a batch - Cloudera Manager restarts the slave roles rack-by-rack in alphabetical order, and within each rack, hosts are restarted in alphabetical order. If you are using the default replication factor of 3, Hadoop tries to keep the replicas on at least 2 different racks. So if you have multiple racks, you can use a higher batch size than the default 1. But you should be aware that using too high batch size also means that fewer slave roles are active at any time during the upgrade, so it can cause temporary performance degradation. If you are using a single rack only, you should only restart *one slave node at a time* to ensure data availability during upgrade.
  - How long should Cloudera Manager wait before starting the next batch.
  - The number of *batch* failures that will cause the entire rolling restart to fail (this is an advanced feature). For example if you have a very large cluster you can use this option to allow failures because if you know that your cluster will be functional even if some worker roles are down.

■ **Note:**

- **HDFS** - If you do not have HDFS high availability configured, a warning appears reminding you that the service will become unavailable during the restart while the NameNode is restarted. Services that depend on that HDFS service will also be disrupted. It is recommended that you restart the DataNodes one at a time—one host per batch, which is the default.
- **HBase** - Administration operations such as any of the following should not be performed during the rolling restart, to avoid leaving the cluster in an inconsistent state:
  - Split
  - Create, disable, enable, or drop table
  - Metadata changes
  - Create, clone, or restore a snapshot. Snapshots rely on the RegionServers being up; otherwise the snapshot will fail.
- **MapReduce** - If you restart the JobTracker, all current jobs will fail.
- **YARN** - If you restart ResourceManager and RM HA is enabled, current jobs continue running; they do not restart or fail. RM HA is supported for CDH 5.2 and higher.
- **ZooKeeper and Flume** - For both ZooKeeper and Flume, the option to restart roles in batches is not available. They are always restarted one by one.

5. Click **Confirm** to start the rolling restart.

### Performing a Cluster-Level Rolling Restart

You can perform a cluster-level rolling restart on demand from the Cloudera Manager Admin Console. A cluster-level rolling restart is also performed as the last step in a rolling upgrade when the cluster is configured with HDFS high availability enabled.

1. If you have not already done so, enable high availability. See [HDFS High Availability](#) on page 211 for instructions. You do not need to enable automatic failover for rolling restart to work, though you can enable it if you wish. Automatic failover does not affect the rolling restart operation.
2. For the cluster you want to restart select **Actions > Rolling Restart**.
3. In the pop-up dialog box, select the services you want to restart. Please review the caveats in the preceding section for the services you elect to have restarted. The services that do not support rolling restart will simply be restarted, and will be unavailable during their restart.
4. If you select an HDFS, HBase, or MapReduce service, you can have their worker roles restarted in batches. You can configure:
  - How many roles should be included in a batch - Cloudera Manager restarts the slave roles rack-by-rack in alphabetical order, and within each rack, hosts are restarted in alphabetical order. If you are using the default replication factor of 3, Hadoop tries to keep the replicas on at least 2 different racks. So if you have multiple racks, you can use a higher batch size than the default 1. But you should be aware that using too high batch size also means that fewer slave roles are active at any time during the upgrade, so it can cause temporary performance degradation. If you are using a single rack only, you should only restart *one slave node at a time* to ensure data availability during upgrade.
  - How long should Cloudera Manager wait before starting the next batch.
  - The number of *batch* failures that will cause the entire rolling restart to fail (this is an advanced feature). For example if you have a very large cluster you can use this option to allow failures because if you know that your cluster will be functional even if some worker roles are down.
5. Click **Confirm** to start the rolling restart. While the restart is in progress, the Command Details page shows the steps for stopping and restarting the services.


## Aborting a Pending Command

**Required Role:** Operator Configurator Cluster Administrator Full Administrator

Commands will time out if they are unable to complete after a period of time.

If necessary, you can abort a pending command. For example, this may become necessary because of a hardware or network failure where a host running a role instance becomes disconnected from the Cloudera Manager Server host. In this case, the Cloudera Manager Server will be unable to connect to the Cloudera Manager Agent on that disconnected host to start or stop the role instance which will prevent the corresponding service from starting or stopping. To work around this, you can abort the command to start or stop the role instance on the disconnected host, and then you can start or stop the service again.

### To abort any pending command:


You can click the indicator () with the blue badge, which shows the number of commands that are currently running in your cluster (if any). This indicator is positioned just to the left of the **Support** link at the right hand side of the navigation bar. Unlike the Commands tab for a role or service, this indicator includes all commands running for all services or roles in the cluster. In the Running Commands window, click **Abort** to abort the pending command. For more information, see [Viewing Running and Recent Commands](#).

### To abort a pending command for a service or role:

1. Navigate to the **Service > Instances** tab for the service where the role instance you want to stop is located. For example, navigate to the **HDFS Service > Instances** tab if you want to abort a pending command for a DataNode.
2. In the list of instances, click the link for role instance where the command is running (for example, the instance that is located on the disconnected host).
3. Go to the **Commands** tab.
4. Find the command in the list of **Running Commands** and click **Abort Command** to abort the running command.

## Deleting Services


**Required Role:** Full Administrator

1. Stop the service. For information on starting and stopping services, see [Starting, Stopping, and Restarting Services](#) on page 33.
2. On the Home page, click  to the right of the service name and select **Delete**.
3. Click **Delete** to confirm the deletion. Deleting a service does *not* clean up the associated [client configurations](#) that have been deployed in the cluster or the user data stored in the cluster. For a given "alternatives path" (for example `/etc/hadoop/conf`) if there exist both "live" client configurations (ones that would be pushed out with deploy client configurations for active services) and ones that have been "orphaned" client configurations (the service they correspond to has been deleted), the orphaned ones will be removed from the alternatives database. In other words, to trigger cleanup of client configurations associated with a deleted service you must create a service to replace it. To remove user data, see [Remove User Data](#).

## Renaming a Service

**Required Role:** Full Administrator

A service is given a name upon installation, and that name is used as an identifier internally. However, Cloudera Manager allows you to provide a display name for a service, and that name will appear in the Cloudera Manager Admin Console instead of the original (internal) name.

1. On the Home page, click  to the right of the service name and select **Rename**.
2. Type the new name.
3. Click **Rename**.

## Configuring CDH and Managed Services

The original service name will still be used internally, and may appear or be required in certain circumstances, such as in log messages or in the API.

The rename action is recorded as an Audit event.

When looking at Audit or Event search results for the renamed service, it is possible that these search results might contain either only the original (internal) name, or both the display name and the original name.

### Configuring Maximum File Descriptors

**Required Role:** **Configurator** **Cluster Administrator** **Full Administrator**

You can setting the maximum file descriptor parameter for all daemon roles. When not specified, the role uses whatever value it inherits from supervisor. When specified, configures soft and hard limits to the configured value.

1. Go to a service.
2. Click the **Configuration** tab.
3. In the Search box, type **rlimit\_fds**.
4. Set the **Maximum Process File Descriptors** property for one or more roles.
5. Click the **Save Changes** button.
6. Restart the affected role instances.

### Managing Roles

When Cloudera Manager configures a service, it configures hosts in your cluster with one or more functions (called roles in Cloudera Manager) that are required for that service. The role determines which Hadoop daemons run on a given host. For example, when Cloudera Manager configures an HDFS service instance it configures one host to run the NameNode role, another host to run as the Secondary NameNode role, another host to run the Balancer role, and some or all of the remaining hosts to run DataNode roles.

Configuration settings are organized in role groups. A **role group** includes a set of configuration properties for a specific group, as well as a list of role instances associated with that role group. Cloudera Manager automatically creates default role groups.

For role types that allow multiple instances on multiple hosts, such as DataNodes, TaskTrackers, RegionServers (and many others), you can create multiple role groups to allow one set of role instances to use different configuration settings than another set of instances of the same role type. In fact, upon initial cluster setup, if you are installing on identical hosts with limited memory, Cloudera Manager will (typically) automatically create two role groups for each worker role — one group for the role instances on hosts with only other worker roles, and a separate group for the instance running on the host that is also hosting master roles.

The HDFS service is an example of this: Cloudera Manager typically creates one role group (DataNode Default Group) for the DataNode role instances running on the worker hosts, and another group (HDFS-1-DATANODE-1) for the DataNode instance running on the host that is also running the master roles such as the NameNode, JobTracker, HBase Master and so on. Typically the configurations for those two classes of hosts will differ in terms of settings such as memory for JVMs.

The Cloudera Manager configuration screens offer two layout options: classic and new. The classic layout is the default; however, on each configuration page you can easily switch between the classic and new layouts using the **Switch to XXX layout** link at the top right of the page. All the configuration procedures described in the Cloudera Manager documentation assume the classic layout.

- **classic** - pages are organized by role group and categories within the role group. For example, to display the JournalNode maximum log size property (JournalNode Max Log Size), select **JournalNode Default Group > Logs**.

**HDFS-1** Status Instances Configuration Commands Audits File Browser Charts ▾

**Configuration**

Q Search ✕

Category	Property	Value
▶ Service-Wide	JournalNode Log Directory	/var/log/hadoop-hdfs default value
▶ Balancer Default Group		
▶ DataNode Default Group		
▶ Failover Controller Default Group	JournalNode Logging Threshold	INFO default value
▶ Gateway Default Group		
▶ HttpFS Default Group		
▼ JournalNode Default Group		
Advanced	JournalNode Max Log Size	200 MIB default value
<b>Logs</b>		
Monitoring		
Performance	JournalNode Maximum Log File Backups	10 default value
Ports and Addresses		
Resource Management		

When a configuration property has been set to a value different from the default, a **Reset to the default value** link displays.

- **new** – pages contain controls that allow you filter configuration properties based on configuration status, category, and group. For example, to display the JournalNode maximum log size property (JournalNode Max Log Size), click the **CATEGORY** > **JournalNode** and **GROUP** > **Logs** filters:

**HDFS-1** Status Instances Configuration Commands Audits File Browser Charts ▾

**Configuration** HDFS-1 on Cluster 1 ▾ Selected Filters: x JournalNode x Logs x Clear All

Reason for change... Save Changes

Filters	Property	Value
▼ SEARCH		
▼ STATUS		
All 4		
Error 0		
Warning 0		
Edited 0		
Non-default 0		
Has Overrides 0		
▼ CATEGORY		
All 31		
HDFS-1 (Service-Wide) 3		
Balancer 0		
DataNode 4		
Gateway 0		
<b>JournalNode 4</b>		
NFS Gateway 4		
NameNode 4		
SecondaryNameNode 4		
Failover Controller 4		
HttpFS 4		
▼ GROUP		
All 50		
Advanced 8		
Checkpointing 0		
Cloudera Navigator 0		
High Availability 0		
<b>Logs 4</b>		
Main 0		

<b>JournalNode Log Directory</b>	JournalNode Default Group /var/log/hadoop-hdfs
<b>JournalNode Logging Threshold</b>	JournalNode Default Group <input type="radio"/> TRACE <input type="radio"/> DEBUG <input checked="" type="radio"/> INFO <input type="radio"/> WARN <input type="radio"/> ERROR <input type="radio"/> FATAL
<b>JournalNode Max Log Size</b>	JournalNode Default Group <input type="text" value="200"/> MIB
<b>JournalNode Maximum Log File Backups</b>	JournalNode Default Group <input type="text" value="10"/>

Display 25 Entries

When a configuration property has been set to a value different from the default, a reset to default value icon displays.

## Gateway Roles

A **gateway** is a special type of role whose sole purpose is to designate a host that should receive a client configuration for a specific service, when the host does not have any roles running on it. Gateway roles enable Cloudera Manager to install and manage client configurations on that host. There is no process associated with a gateway role, and its status will always be Stopped. You can configure gateway roles for HBase, HDFS, Hive, MapReduce, Solr, Sqoop 1 Client, and YARN.

Role Instances

Adding a Role Instance

Required Role: **Cluster Administrator** **Full Administrator**

After creating services, you can add role instances to the services. For example, after initial installation in which you created the HDFS service, you can add a DataNode role instance to a host where one was not previously running. Upon upgrading a cluster to a new version of CDH you might want to create a role instance for a role added in the new version. For example, in CDH 5 Impala has the Impala Llama ApplicationMaster role, which must be added after you upgrade a CDH 4 cluster to CDH 5.

- 1. Go to the service for which you want to add a role instance. For example, to add a DataNode role instance, go to the HDFS service.
- 2. Click the **Instances** tab.
- 3. Click the **Add Role Instances** button.
- 4. Customize the assignment of role instances to hosts. The wizard evaluates the hardware configurations of the hosts to determine the best hosts for each role. The wizard assigns all worker roles to the same set of hosts to which the HDFS DataNode role is assigned. These assignments are typically acceptable, but you can reassign them if necessary.

Click a field below a role to display a dialog containing a list of hosts. If you click a field containing multiple hosts, you can also select **All Hosts** to assign the role to all hosts or **Custom** to display the pageable hosts dialog.

The following shortcuts for specifying hostname patterns are supported:

- Range of hostnames (without the domain portion)

Range Definition	Matching Hosts
10.1.1.[1-4]	10.1.1.1, 10.1.1.2, 10.1.1.3, 10.1.1.4
host[1-3].company.com	host1.company.com, host2.company.com, host3.company.com
host[07-10].company.com	host07.company.com, host08.company.com, host09.company.com, host10.company.com

- IP addresses
- Rack name

Click the **View By Host** button for an overview of the role assignment by hostname ranges.

- 5. Click **Continue**.
- 6. In the Review Changes page, review the configuration changes to be applied. Confirm the settings entered for file system paths. The file paths required vary based on the services to be installed. For example, you might confirm the NameNode Data Directory and the DataNode Data Directory for HDFS. Click **Continue**. The wizard finishes by performing any actions necessary to prepare the cluster for the new role instances. For example, new DataNodes are added to the NameNode `dfs_hosts_allow.txt` file. The new role instance is configured with the default role group for its role type, even if there are multiple role groups for the role type. If you want to use a different role group, follow the instructions in [Managing Role Groups](#) on page 43 for moving role instances to a different role group. The new role instances are not started automatically.

Starting, Stopping, and Restarting Role Instances

Required Role: **Operator** **Configurator** **Cluster Administrator** **Full Administrator**

If the host for the role instance is currently decommissioned, you will not be able to start the role until the host has been recommissioned.



1. Go to the service that contains the role instances to start, stop, or restart.
2. Click the **Instances** tab.
3. Check the checkboxes next to the role instances to start, stop, or restart (such as a DataNode instance).
4. Select **Actions for Selected** > **Start**, **Stop**, or **Restart**, and then click **Start**, **Stop**, or **Restart** again to start the process. When you see a **Finished** status, the process has finished.

Also see [Rolling Restart](#) on page 34.

### Decommissioning Role Instances



**Required Role:** **Operator** **Configurator** **Cluster Administrator** **Full Administrator**

You can remove a role instance such as a DataNode from a cluster while the cluster is running by decommissioning the role instance. When you decommission a role instance, Cloudera Manager performs a procedure so that you can safely retire a host without losing data. Role decommissioning applies to HDFS DataNode, MapReduce TaskTracker, YARN NodeManager, and HBase RegionServer roles.

You cannot decommission a DataNode or a host with a DataNode if the number of DataNodes equals the replication factor (which by default is three) of any file stored in HDFS. For example, if the replication factor of any file is three, and you have three DataNodes, you cannot decommission a DataNode or a host with a DataNode.

A role will be decommissioned if its host is decommissioned. See [Decommissioning and Recommissioning Hosts](#) on page 107 for more details.

To decommission role instances:

1. If you are decommissioning DataNodes, perform the steps in [Tuning HDFS Prior to Decommissioning DataNodes](#) on page 108.
2. Click the service instance that contains the role instance you want to decommission.
3. Click the **Instances** tab.
4. Check the checkboxes next to the role instances to decommission.
5. Select **Actions for Selected** > **Decommission**, and then click **Decommission** again to start the process. While decommissioning is in progress, the role instance displays the  icon. If one role instance of a service is decommissioned, the DECOMMISSIONED facet displays in the Filters on the Instances page and the  icon displays on the role instance's page. When you see a **Finished** status, the decommissioning process has finished.

### Recommissioning Role Instances

**Required Role:** **Operator** **Configurator** **Cluster Administrator** **Full Administrator**

1. Click the service instance that contains the role instance you want to recommission.
2. Click the **Instances** tab.
3. Check the checkboxes next to the decommissioned role instances to recommission.
4. Select **Actions for Selected** > **Recommission**, and then click **Recommission** again to start the process. When you see a **Finished** status, the recommissioning process has finished.

### Deleting Role Instances

**Required Role:** **Cluster Administrator** **Full Administrator**

1. Click the service instance that contains the role instance you want to delete. For example, if you want to delete a DataNode role instance, click an HDFS service instance.
2. Click the **Instances** tab.
3. Check the checkboxes next to the role instances you want to delete.
4. If the role instance is running, select **Actions for Selected** > **Stop** and click **Stop** to confirm the action.

5. Select **Actions for Selected** > **Delete**. Click **Delete** to confirm the deletion.

- **Note:** Deleting a role instance does not clean up the associated client configurations that have been deployed in the cluster.

### Configuring Roles to Use a Custom Garbage Collection Parameter

Every Java-based role in Cloudera Manager has a configuration setting called **Java Configuration Options for role** where you can enter command line options. Commonly, garbage collection flags or extra debugging flags would be passed here. To find the appropriate configuration setting, select the service you want to modify in the Cloudera Manager Admin Console, then use the Search box to search for Java Configuration Options.

You can add configuration options for all instances of a given role by making this configuration change at the service level. For example, to modify the setting for all DataNodes, select the HDFS service, then modify the **Java Configuration Options for DataNode** setting.

To modify a configuration option for a given instance of a role, select the service, then select the particular role instance (for example, a specific DataNode). The configuration settings you modify will apply to the selected role instance only.

For detailed instructions see [Modifying Configuration Properties](#) on page 8.

### Role Groups

**Required Role:** **Configurator** **Cluster Administrator** **Full Administrator**

A **role group** is a set of configuration properties for a role type, as well as a list of role instances associated with that group. Cloudera Manager automatically creates a default role group named **Role Type Default Group** for each role type. Each role instance can be associated with only a single role group.

Role groups provide two types of properties: those that affect the configuration of the service itself and those that affect monitoring of the service, if applicable (the **Monitoring** subcategory). (Not all services have monitoring properties). For more information about monitoring properties see [Configuring Monitoring Settings](#).

When you run the installation or upgrade wizard, Cloudera Manager automatically creates the appropriate configurations for the default role groups it adds. It may also create additional role groups for a given role type, if necessary. For example, if you have a DataNode role on the same host as the NameNode, it may require a slightly different configuration than DataNode roles running on other hosts. Therefore, Cloudera Manager will create a separate role group for the DataNode role that is running on the NameNode host, and use the default DataNode configuration for the DataNode roles running on other hosts.

You can modify the settings of the default role group, or you can create new role groups and associate role instances to whichever role group is most appropriate. This simplifies the management of role configurations when one group of role instances may require different settings than another group of instances of the same role type—for example, due to differences in the hardware the roles run on. You modify the configuration for any of the service's role groups through the Configuration tab for the service. You can also [override](#) the settings inherited from a role group for a role instance.

If there are multiple role groups for a role type, you can move role instances from one group to another. When you move a role instance to a different group, it inherits the configuration settings for its new group.

### Creating a Role Group

1. Go to a service status page.
2. Click the **Instances** or **Configuration** tab.
3. Click **Role Groups**.
4. Click **Create new group...**
5. Provide a name for the group.
6. Select the role type for the group. You can select role types that allow multiple instances and that exist for the service you have selected.



7. In the **Copy From** field, select the source of the basic configuration information for the role group:

- An existing role group of the appropriate type.
- **None....** The role group is set up with generic default values that are *not* the same as the values Cloudera Manager sets in the default role group, as Cloudera Manager specifically sets the appropriate configuration properties for the services and roles it installs. After you create the group you must [edit the configuration](#) to set missing properties (for example the TaskTracker Local Data Directory List property, which is not populated if you select None) and clear other validation warnings and errors.

### Managing Role Groups

You can rename or delete existing role groups, and move roles of the same role type from one group to another.

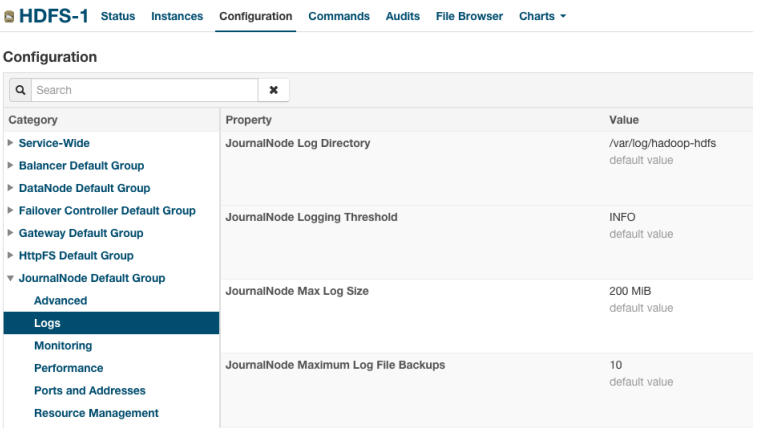
1. Go to a service status page.
2. Click the **Instances** or **Configuration** tab.
3. Click **Role Groups**.
4. Click the group you want to manage. Role instances assigned to the role group are listed.
5. Perform the appropriate procedure for the action:

Action	Procedure
<b>Rename</b>	<ol style="list-style-type: none"> <li>1. Click the role group name, click  next to the name on the right and click <b>Rename</b>.</li> <li>2. Specify the new name and click <b>Rename</b>.</li> </ol>
<b>Delete</b>	<p>You cannot delete any of the default groups. The group must first be empty; if you want to delete a group you've created, you must move any role instances to a different role group.</p> <ol style="list-style-type: none"> <li>1. Click the role group name.</li> <li>2. Click  next to the role group name on the right, select <b>Delete</b>, and confirm by clicking <b>Delete</b>. Deleting a role group removes it from <a href="#">host templates</a>.</li> </ol>
<b>Move</b>	<ol style="list-style-type: none"> <li>1. Select the role instance(s) to move.</li> <li>2. Select <b>Actions for Selected</b> &gt; <b>Move To Different Role Group....</b></li> <li>3. In the pop-up that appears, select the target role group and click <b>Move</b>.</li> </ol>

### Managing Individual Services

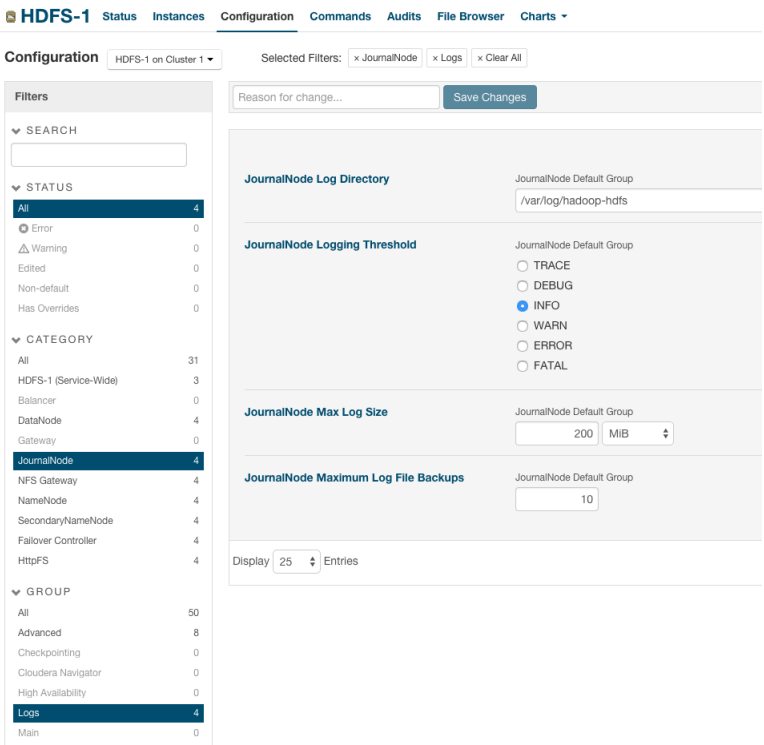
The Cloudera Manager configuration screens offer two layout options: classic and new. The classic layout is the default; however, on each configuration page you can easily switch between the classic and new layouts using the **Switch to XXX layout** link at the top right of the page. All the configuration procedures described in the Cloudera Manager documentation assume the classic layout.

- **classic** - pages are organized by role group and categories within the role group. For example, to display the JournalNode maximum log size property (JournalNode Max Log Size), select **JournalNode Default Group** > **Logs**.



When a configuration property has been set to a value different from the default, a **Reset to the default value** link displays.

- **new** – pages contain controls that allow you filter configuration properties based on configuration status, category, and group. For example, to display the JournalNode maximum log size property (JournalNode Max Log Size), click the **CATEGORY** > **JournalNode** and **GROUP** > **Logs** filters:



When a configuration property has been set to a value different from the default, a reset to default value icon displays.

The following sections cover the configuration and management of individual CDH and other services, that have specific and unique requirements or options.


### The Flume Service

The Flume packages are installed by the Installation wizard, but the service is not created. To add the Flume service, follow these steps:

1. [Add the Flume service.](#)
2. [Configure the Flume agents.](#)
3. Start the Flume service, which will start all the Flume agents.

### Adding a Flume Service

**Required Role:** Cluster Administrator Full Administrator

1. On the Home page, click  to the right of the cluster name and select **Add a Service**. A list of service types display. You can add one type of service at a time.
2. Select the **Flume** service and click **Continue**.
3. Select the radio button next to the services on which the new service should depend and click **Continue**.
4. Customize the assignment of role instances to hosts. The wizard evaluates the hardware configurations of the hosts to determine the best hosts for each role. The wizard assigns all worker roles to the same set of hosts to which the HDFS DataNode role is assigned. These assignments are typically acceptable, but you can reassign them if necessary.

Click a field below a role to display a dialog containing a list of hosts. If you click a field containing multiple hosts, you can also select **All Hosts** to assign the role to all hosts or **Custom** to display the pageable hosts dialog.

The following shortcuts for specifying hostname patterns are supported:

- Range of hostnames (without the domain portion)

Range Definition	Matching Hosts
10.1.1.[1-4]	10.1.1.1, 10.1.1.2, 10.1.1.3, 10.1.1.4
host[1-3].company.com	host1.company.com, host2.company.com, host3.company.com
host[07-10].company.com	host07.company.com, host08.company.com, host09.company.com, host10.company.com

- IP addresses
- Rack name

Click the **View By Host** button for an overview of the role assignment by hostname ranges.

### Configuring the Flume Agents

**Required Role:** Configurator Cluster Administrator Full Administrator

After you create a Flume service, you must first configure the agents before you start them. For detailed information about Flume agent configuration, see the [Flume User Guide](#).

The default Flume agent configuration provided in the **Configuration File** property of the **Agent** default role group is a configuration for a single agent in a single tier; you should replace this with your own configuration. When you add new agent roles, they are placed (initially) in the **Agent** default role group.

Agents that share the same configuration should be members of the same agent role group. You can create new [role groups](#) and can move agents between them. If your Flume configuration has multiple tiers, you must *create an agent role group for each tier*, and move each agent to be a member of the appropriate role group for their tier.

A Flume agent role group **Configuration File** property can contain the configuration for multiple agents, since each configuration property is prefixed by the agent name. You can [set the agents' names](#) using configuration overrides to change the name of a specific agent without changing its role group membership. Different agents can have the same name — agent names do not have to be unique.

## Configuring CDH and Managed Services

1. Go to the Flume service.
2. Click the **Configuration** tab.
3. Select the **Agent** default role group in the left hand column. The settings you make here apply to the default role group, and thus will apply to all agent instances unless you associate those instances with a different role group, or override them for specific agents.
4. Set the **Agent Name** property to the name of the agent (or one of the agents) defined in the `flume.conf` configuration file. The agent name can be comprised of letters, numbers, and the underscore character. You can specify only one agent name here — the name you specify will be used as the default for all Flume agent instances, unless you override the name for specific agents. You can have multiple agents with the same name — they will share the configuration specified in on the configuration file.
5. Copy the contents of the `flume.conf` file, in its entirety, into the **Configuration File** property. Unless overridden for specific agent instances, this property applies to all agents in the group. You can provide multiple agent configurations in this file and use agent name overrides to specify which configuration to use for each agent.

■ **Important:** The name-value property pairs in the **Configuration File** property *must* include an equal sign (=). For example, `tier1.channels.channel1.capacity = 10000`.

### Setting a Flume Agent Name

**Required Role:** **Configurator** **Cluster Administrator** **Full Administrator**

If you have specified multiple agent configurations in a Flume agent role group **Configuration File** property, you can set the agent name for an agent that uses a different configuration. Overriding the agent name will point the agent to the appropriate properties specified in the agent configuration.

1. Go to the Flume service.
2. Click the **Configuration** tab.
3. Select an agent role group in the left hand column.
4. To override the agent name for one or more instances within the group, move your cursor over the Value column of the **Agent Name** property row, and click **Override Instances**.
5. Select the agent role instances to override.
6. In the field **Change value of selected instances to:** select **Other**. (You can use the **Inherited Value** setting to return to the role group value.)
7. In the field that appears, type the name for the selected agent instance.
8. Click **Apply** to have your change take effect.

### Using Flume with HDFS or HBase Sinks

If you want to use Flume with HDFS or HBase sinks, you can add a dependency to that service from the Flume configuration page. This will automatically add the correct client configurations to the Flume agent's classpath.

■ **Note:** If you are using Flume with HBase, make sure that the `/etc/zookeeper/conf/zoo.cfg` file either does not exist on the host of the Flume agent that is using an HBase sink, or that it contains the correct ZooKeeper quorum.

### Using Flume with Solr Sinks

Cloudera Manager provides a set of configuration settings under the Flume service to configure the Flume Morphline Solr Sink. See [Configuring the Flume Morphline Solr Sink for Use with the Solr Service](#) on page 91 for detailed instructions.

### Updating Flume Agent Configurations

**Required Role:** **Full Administrator**

If you modify the **Configuration File** property after you have started the Flume service, update the configuration across Flume agents as follows:

1. Go to the Flume service.
2. Select **Actions** > **Update Config**.

## The HBase Service

Cloudera Manager requires certain additional steps to set up and configure the HBase service.

### Creating the HBase Root Directory

**Required Role:** Cluster Administrator Full Administrator

When adding the HBase service, the **Add Service** wizard automatically creates a root directory for HBase in HDFS. If you quit the **Add Service** wizard or it does not finish, you can create the root directory outside the wizard by doing these steps:

1. Choose **Create Root Directory** from the **Actions** menu in the **HBase** > **Status** tab.
2. Click **Create Root Directory** again to confirm.

### Graceful Shutdown

**Required Role:** Operator Configurator Cluster Administrator Full Administrator

A graceful shutdown of an HBase RegionServer allows the regions hosted by that RegionServer to be moved to other RegionServers before stopping the RegionServer. Cloudera Manager provides the following configuration options to perform a graceful shutdown of either an HBase RegionServer or the entire service.

#### Gracefully Shutting Down an HBase RegionServer

1. Go to the HBase service.
2. Click the **Instances** tab.
3. From the list of Role Instances, select the RegionServer you want to shut down gracefully.
4. Select **Actions for Selected** > **Decommission (Graceful Stop)**.

- **Note:** To simply shut down a RegionServer without moving regions off it, select **Actions for Selected** > **Stop**. This is only recommended if graceful shutdown isn't working for some reason, since this can cause regions to become temporarily unavailable.

#### Gracefully Shutting Down the HBase Service

1. Go to the HBase service.
2. Select **Actions** > **Stop**. This tries to perform an HBase Master-driven graceful shutdown for the length of the configured Graceful Shutdown Timeout (three minutes by default), after which it abruptly shuts down the whole service.

#### Configuring the Graceful Shutdown Timeout Property

**Required Role:** Configurator Cluster Administrator Full Administrator

This timeout only affects a graceful shutdown of the entire HBase service, not individual RegionServers. Therefore, if you have a large cluster with many RegionServers, you should strongly consider increasing the timeout from its default of 180 seconds.

1. Go to the HBase service.
2. Click the **Configuration** tab.
3. Under the **Service-Wide** category, scroll down to the **Graceful Shutdown Timeout** property and click on the current Value to change it.

## Configuring CDH and Managed Services

4. Click **Save Changes** to save this setting.

### Adding the HBase Thrift Server

**Required Role:** **Cluster Administrator** **Full Administrator**

The Thrift Server role is not added by default when you install HBase. To add the Thrift Server role:

1. Go to the HBase service.
2. Click the **Instances** tab.
3. Click the **Add Role Instances** button.
4. Select the host(s) where you want to add the Thrift Server role (you only need one for Hue) and click **Continue**.  
The Thrift Server role should appear in the instances list for the HBase server.
5. Select the Thrift Server role instance.
6. Select **Actions for Selected** > **Start**.

### Enabling HBase Indexing

**Required Role:** **Configurator** **Cluster Administrator** **Full Administrator**

HBase indexing is dependent on the [Key-Value Store Indexer service](#). The Key-Value Store Indexer service uses the [Lily HBase Indexer Service](#) to index the stream of records being added to HBase tables. Indexing allows you to query data stored in HBase with the [Solr service](#).

1. Go to the HBase service.
2. Click the **Configuration** tab.
3. Select the **Backup** category.
4. Check the properties for **Enable Replication** and **Enable Indexing**, and click **Save Changes**.

### Adding a Custom Coprocessor

**Required Role:** **Configurator** **Cluster Administrator** **Full Administrator**

The HBase coprocessor framework provides a way to extend HBase with custom functionality. The following properties can be configured for HBase custom coprocessors from the **Configuration** tab for the HBase service:

- Click **Service Wide**.
- Choose *one* of these groups: **Master Default Group** or **RegionServer Default Group**.
- You can configure the values of the following properties:
  - **HBase Coprocessor Abort on Error**
  - **HBase Coprocessor Master Classes** or **HBase Coprocessor Region Classes**
- Click **Save Changes** to commit the changes.

### Enabling Hedged Reads on HBase

1. Go to the HBase service.
2. Click the **Configuration** tab.
3. Expand the **Service-Wide** category.
4. Select **Performance**.
5. Configure the **HDFS Hedged Read Threadpool Size** and **HDFS Hedged Read Delay Threshold** properties. The descriptions for each of these properties on the configuration pages provide more information.
6. Click **Save Changes** to commit the changes.

### Hedged Reads

Hadoop 2.4 introduced a new feature called *hedged reads*, in [HDFS-5776](#). If a read from a block is slow, the HDFS client starts up another parallel, 'hedged' read against a different block replica. The result of whichever read returns first is used, and the outstanding read is cancelled. This feature helps in situations where a read



occasionally takes a long time rather than when there is a systemic problem. Hedged reads can be enabled for HBase when the HFiles are stored in HDFS. This feature is disabled by default.

### Enabling Hedged Reads for HBase Using Cloudera Manager

Required Role: **Configurator** **Cluster Administrator** **Full Administrator**

1. Go to the HBase service.
2. Click the **Configuration** tab.
3. Expand the **Service-Wide** category.
4. Select **Performance**.
5. Configure the **HDFS Hedged Read Threadpool Size** and **HDFS Hedged Read Delay Threshold** properties. The descriptions for each of these properties on the configuration pages provide more information.
6. Click **Save Changes** to commit the changes.

### Enabling Hedged Reads for HBase Using the Command Line

■ **Important:**

- If you use Cloudera Manager, do not use these command-line instructions.
- This information applies specifically to CDH 5.3.x. If you use an earlier version of CDH, see the documentation for that version located at [Cloudera Documentation](#).

To enable hedged reads for HBase, edit the `hbase-site.xml` file on each server. Set `dfs.client.hedged.read.threadpool.size` to the number of threads to dedicate to running hedged threads, and set the `dfs.client.hedged.read.threshold.millis` configuration property to the number of milliseconds to wait before starting a second read against a different block replica. Set `dfs.client.hedged.read.threadpool.size` to 0 or remove it from the configuration to disable the feature. After changing these properties, restart your cluster.

The following is an example configuration for hedged reads for HBase.

```
<property>
  <name>dfs.client.hedged.read.threadpool.size</name>
  <value>20</value>  <!-- 20 threads -->
</property>
<property>
  <name>dfs.client.hedged.read.threshold.millis</name>
  <value>10</value>  <!-- 10 milliseconds -->
</property>
```

### Monitoring the Performance of Hedged Reads

You can monitor the performance of hedged reads using the following metrics emitted by Hadoop when hedged reads are enabled.

- **hedgedReadOps** - the number of hedged reads that have occurred
- **hedgedReadOpsWin** - the number of times the hedged read returned faster than the original read

## The HDFS Service

The section contains configuration tasks for the HDFS service. For information on configuring HDFS for high availability, see [HDFS High Availability](#) on page 211.

### Federated Nameservices

Required Role: **Cluster Administrator** **Full Administrator**

Cloudera Manager supports the configuration of multiple nameservices managing separate HDFS namespaces, all of which share the storage available on the set of DataNodes. These nameservices are federated, meaning

## Configuring CDH and Managed Services

each nameservice is independent and does not require coordination with other nameservices. See [HDFS Federation](#) for more information.

It is simplest to add a second nameservice if high availability is already enabled. The process of enabling high availability creates a nameservice as part of the enable high availability workflow.

- **Important:** Configuring a new nameservice will shut down the services that depend upon HDFS. Once the new nameservice has been started, the services that depend upon HDFS must be restarted, and the client configurations must be redeployed. (This can be done as part of the **Add Nameservice** workflow, as an option.)

### *Converting a non-Federated HDFS Service to a Federated HDFS Service*

You must have one nameservice in place before you can add a second (or additional) nameservices. Follow the instructions below to convert your current NameNode/SecondaryNameNode setup to a Federated Setup with a single nameservice.

1. Go to the HDFS service.
2. Click the **Configuration** tab.
3. Search for "nameservice". This will show you the nameservice properties for your NameNode and SecondaryNameNode.
4. In the **NameNode Nameservice** field, type a name for the nameservice. The name must be unique and can contain only alphanumeric characters.
5. In the **Mountpoints** field, change the mount point from "/" to a list of mount points that are in the namespace that this nameservice will manage. (You can enter this as a comma-separated list — for example, `"/hbase, /tmp, /user"` or by clicking the plus icon to add each mount point in its own field.) You can determine the list of mount points by running the command `hadoop fs -ls /` from the CLI on the NameNode host.
6. In the **SecondaryNameNode Nameservice** field, type the name of the nameservice. This must be the same as you provided for the **NameNode Nameservice** property.
7. Save your changes.
8. Click the **Instances** tab. You should now see the **Federation and High Availability** section with your nameservice listed.
9. You can use the **Edit** command under the **Actions** menu to edit the list of mount points for this nameservice. In the **Mountpoints** field, change the mount point from "/" to a list of mount points that are in the namespace that this nameservice will manage.

### *Adding a Nameservice*

The instructions below for adding a nameservice assume that a nameservice is already set up. The first nameservice can be set up either by converting a simple HDFS service as described above (see [Converting a non-Federated HDFS Service to a Federated HDFS Service](#) on page 50 or by enabling [HDFS High Availability](#) on page 211.

1. Go to the HDFS service.
2. Click the **Instances** tab. At the top of this page you should see the **Federation and High Availability** section.

- **Note:** If this section does not appear, it means you do not have any nameservices configured. You must have one nameservice already configured in order to add a nameservice. You can either enable high availability, which will create a nameservice, or you can convert your existing HDFS service to be federated.

3. Click the **Add Nameservice** button.
  - a. Enter a name for the new nameservice. The name must be unique and can contain only alphanumeric characters.
  - b. Enter at least one mount point for the nameservice. This defines the portion of HDFS that will be managed under the new nameservice. (Click the **+** to the right of the Mount Point field to add a new mount point). You cannot use "/" as a mount point; you must specify HDFS directories by name.

- The mount points must be unique for this nameservice; you cannot specify any of the same mount points you have used for other nameservices.
  - You can specify mount points that do not yet exist, and create the corresponding directories in a later step in this procedure.
  - If you want to use a mount point previously associated with another nameservice you must first remove that mount point from that service. You can do this using the **Edit** command from the **Actions** menu for that nameservice, and later add the mount point to the new nameservice.
  - After you have brought up the new nameservice, you must create the directories that correspond with the mount points you specified in the new namespace.
  - If a mount point corresponds to a directory that formerly was under a different nameservice, you must also move any contents of that directory, if appropriate as described in [step 8](#).
  - If an HBase service is set to depend on the federated HDFS service, edit the mount points of the existing nameservice to reference:
    - HBase root directory (default `/hbase`)
    - MapReduce system directory (default `/tmp/mapred/system`)
    - MapReduce JobTracker staging root directory (default value `/user`).
- c. If you want to configure high availability for the nameservice, leave the **Highly Available** checkbox checked.
- d. Click **Continue**.
4. Select the hosts on which the new NameNode and Secondary NameNodes will be created. (These must be hosts that are not already running other NameNode or SecondaryNameNode instances, and their `/dfs/nn` and `/dfs/snn` directories should be empty if they exist. Click **Continue**).
  5. Enter or confirm the directory property values (these will differ depending on whether you are enabling high availability for this nameservice, or not).
  6. Uncheck the **Start Dependent Services** checkbox if you need to create directories or move data onto the new nameservice. Leave this checked if you want the workflow to restart services and redeploy the client configurations as the last steps in the workflow.
  7. Click **Continue**. If the process finished successfully, click **Finish**. You should now see your new nameservice in the **Federation and High Availability** section in the **Instances** tab of the HDFS service.
  8. Create the directories you want under the new nameservice using the CLI:
    - a. To create a directory in the new namespace, use the command `hadoop fs -mkdir /nameservices/nameservice/directory` where *nameservice* is the new nameservice you just created and *directory* is the directory that corresponds to a mount point you specified.
    - b. To move data from one nameservice to another, use `distcp` or manual export/import. `dfs -cp` and `dfs -mv` will not work.
    - c. Verify that the directories and data are where you expect them to be.
  9. Restart the dependent services.

▪ **Note:** The monitoring configurations at the HDFS level apply to *all* nameservices. So if you have two nameservices, it is not possible to disable a check on one but not the other. Likewise, it's not possible to have different thresholds for events for the two nameservices.

### *Nameservice and Quorum-based Storage*

With Quorum-based Storage, JournalNodes are shared across nameservices. So, if JournalNodes are present in an HDFS service, all nameservices will have Quorum-based Storage enabled. To override this:

- The `dfs.namenode.shared.edits.dir` configuration of the two NameNodes of a high availability nameservice should be configured to include the value of the `dfs.namenode.name.dirs` setting, or
- The `dfs.namenode.edits.dir` configuration of the one NameNode of a non-high availability nameservice should be configured to include the value of the `dfs.namenode.name.dirs` setting.

## Configuring CDH and Managed Services

### NameNodes

Required Role: **Cluster Administrator** **Full Administrator**

#### *Formatting the NameNode and Creating the /tmp Directory*

When you add an HDFS service, the wizard automatically formats the NameNode and creates the `/tmp` directory on HDFS. If you quit the wizard or it does not finish, you can format the NameNode and create the `/tmp` directory outside the wizard by doing these steps:

1. Stop the HDFS service if it is running. See [Starting, Stopping, and Restarting Services](#) on page 33.
2. In the **HDFS > Instances** tab, click the NameNode role instance.
3. Select **Actions > Format**.
4. In the **HDFS > Instances** tab, click the NameNode role instance.
5. Select **Actions > Create /tmp Directory**.
6. Start the HDFS service.

#### *Moving a NameNode to a Different Host*

If the NameNode is not highly available and the NameNode host has hardware problems, you can move the NameNode to another host as follows:

1. If the host to which you want to move the NameNode is not in the cluster, follow the instructions in [Adding a Host to the Cluster](#) on page 101 to add the host.
2. [Stop all cluster services](#).
3. Make a backup of the `dfs.name.dir` directories on the existing NameNode host. Make sure you back up the `fsimage` and `edits` files. They should be the same across all of the directories specified by the `dfs.name.dir` property.
4. Copy the files you backed up from `dfs.name.dir` directories on the old NameNode host to the host where you want to run the NameNode.
5. Go to the HDFS service.
6. Click the **Instances** tab.
7. Select the checkbox next to the NameNode role instance and then click the **Delete** button. Click **Delete** again to confirm.
8. In the **Review configuration changes** page that appears, click **Skip**.
9. Click **Add Role Instances** to add a NameNode role instance.
10. Select the host where you want to run the NameNode and then click **Continue**.
11. Specify the location of the `dfs.name.dir` directories where you copied the data on the new host, and then click **Accept Changes**.
12. [Start cluster services](#). After the HDFS service has started, Cloudera Manager distributes the new configuration files to the DataNodes, which will be configured with the IP address of the new NameNode host.
13. Go to the HDFS service. The NameNode, Secondary NameNode, and DataNode roles should each show a process state of **Started**, and the HDFS service should show a status of **Good**.

### DataNodes

Required Role: **Operator** **Configurator** **Cluster Administrator** **Full Administrator**

#### *How NameNode Manages Blocks on a Failed DataNode*

After a period without any heartbeats (which by default is 10.5 minutes), a DataNode is assumed to be failed. The following describes how the NameNode manages block replication in such cases.

1. NameNode determines which blocks were on the failed DataNode.
2. NameNode locates other DataNodes with copies of these blocks.
3. The DataNodes with block copies are instructed to copy those blocks to other DataNodes to maintain the configured replication factor.

- Follow the procedure in [Replacing a Disk on a DataNode Host](#) on page 53 to bring a repaired DataNode back online.

### Replacing a Disk on a DataNode Host

If one of your DataNode hosts experiences a disk failure, follow this process to replace the disk:

- Stop managed services.
- [Decommission](#) the DataNode role instance.
- Replace the failed disk.
- Recommission the DataNode role instance.
- Run the HDFS `fsck` utility to validate the health of HDFS. The utility normally reports over-replicated blocks immediately after a DataNode is reintroduced to the cluster, which is automatically corrected over time.
- Start managed services.

### Backing Up HDFS Metadata

Required Role: [Cluster Administrator](#) [Full Administrator](#)

- Note:** Cloudera recommends backing up HDFS metadata before a major upgrade.

- Stop the cluster. It is particularly important that the NameNode role process is not running so that you can make a consistent backup.
- Go to the HDFS service.
- Click the **Configuration** tab.
- In the Search field, search for "NameNode Data Directories". This locates the NameNode Data Directories property.
- From the command line on the NameNode host, back up the directory listed in the NameNode Data Directories property. If more than one is listed, then you only need to make a backup of one directory, since each directory is a complete copy. For example, if the data directory is `/mnt/hadoop/hdfs/name`, do the following as root:

```
# cd /mnt/hadoop/hdfs/name
# tar -cvf /root/nn_backup_data.tar .
```

You should see output like this:

```
./
./current/
./current/fsimage
./current/fstime
./current/VERSION
./current/edits
./image/
./image/fsimage
```

- Warning:** If you see a file containing the word *lock*, the NameNode is probably still running. Repeat the preceding steps, starting by shutting down the CDH services.

### Configuring HDFS Trash

Required Role: [Configurator](#) [Cluster Administrator](#) [Full Administrator](#)

The Hadoop trash feature helps prevent accidental deletion of files and directories. If trash is enabled and a file or directory is deleted using the Hadoop shell, the file is moved to the `.Trash` directory in the user's home directory instead of being deleted. Deleted files are initially moved to the `Current` sub-directory of the `.Trash` directory, and their original path is preserved. Files in `.Trash` are permanently removed after a user-configurable time interval. The interval setting also enables trash checkpointing, where the `Current` directory is periodically renamed using a timestamp. Files and directories in the trash can be restored simply by moving them to a location outside the `.Trash` directory.

## Configuring CDH and Managed Services

The Hadoop Trash feature applies to all services in the cluster that use the trash deletion shell function.

### Enabling and Disabling Trash

1. Go to the HDFS service.
2. Click the **Configuration** tab.
3. Click the **Gateway Default Group** category.
4. Check or uncheck the **Use Trash** checkbox.
5. Click the **Save Changes** button.
6. Restart the cluster and deploy the cluster Client Configuration.

### Setting the Trash Interval

1. Go to the HDFS service.
2. Click the **Configuration** tab.
3. Click the **NameNode Default Group** category.
4. Specify the **Filesystem Trash Interval** property, which controls the number of minutes after which a trash checkpoint directory is deleted and the number of minutes between trash checkpoints. For example, to enable trash so that deleted files are deleted after 24 hours, set the value of the **Filesystem Trash Interval** property to 1440.

- **Note:** The trash interval is measured from the point at which the files are moved to trash, not from the last time the files were modified.

5. Click the **Save Changes** button.
6. Restart all NameNodes.

### The HDFS Balancer

**Required Role:** **Cluster Administrator** **Full Administrator**

HDFS data might not always be placed uniformly across DataNodes. One common reason is addition of new DataNodes to an existing cluster. HDFS provides a balancer utility that analyzes block placement and balances data across the DataNodes. It moves blocks until the cluster is deemed to be balanced, which means that the utilization of every DataNode (ratio of used space on the node to total capacity of the node) differs from the utilization of the cluster (ratio of used space on the cluster to total capacity of the cluster) by no more than a given threshold percentage. The balancer does not balance between individual volumes on a single DataNode.

In Cloudera Manager, the HDFS balancer utility is implemented by the Balancer role. The Balancer role usually shows a health of **None** on the HDFS Instances tab because it does not run continuously.

- **Note:** The Balancer role is normally added (by default) when the HDFS service is installed. If it has not been added, you must add a Balancer role instance in order to rebalance HDFS and to see the **Rebalance** action.

### Running the Balancer

1. Go to the HDFS service.
2. Select **Actions** > **Rebalance**.
3. Click **Rebalance** that appears in the next screen to confirm. If you see a **Finished** status, the Balancer ran successfully.

### Configuring the Balancer Threshold

The Balancer has a default threshold of 10%, which ensures that disk usage on each DataNode differs from the overall usage in the cluster by no more than 10%. For example, if overall usage across all the DataNodes in the

cluster is 40% of the cluster's total disk-storage capacity, the script ensures that DataNode disk usage is between 30% and 50% of the DataNode disk-storage capacity. To change the threshold:

1. Go to the HDFS service.
2. Click the **Configuration** tab.
3. Expand the **Balancer Default Group** category.
4. Set the **Rebalancing Threshold** property.
5. Click **Save Changes** to commit the changes.

### Enabling WebHDFS

**Required Role:** **Configurator** **Cluster Administrator** **Full Administrator**

To enable WebHDFS, proceed as follows:

1. Select the HDFS service.
2. Click the **Configuration** tab.
3. Expand the **Service-Wide** category and click the **Enable WebHDFS** checkbox.
4. Click the **Save Changes** button.
5. Restart the HDFS service.

You can find a full explanation of the WebHDFS API in the [WebHDFS API documentation](#).

### Adding HttpFS

**Required Role:** **Cluster Administrator** **Full Administrator**

Apache Hadoop HttpFS is a service that provides HTTP access to HDFS.

HttpFS has a REST HTTP API supporting all HDFS filesystem operations (both read and write).

Common HttpFS use cases are:

- Read and write data in HDFS using HTTP utilities (such as `curl` or `wget`) and HTTP libraries from languages other than Java (such as Perl).
- Transfer data between HDFS clusters running different versions of Hadoop (overcoming RPC versioning issues), for example using Hadoop DistCp.
- Read and write data in HDFS in a cluster behind a firewall. (The HttpFS server acts as a gateway and is the only system that is allowed to send and receive data through the firewall).

HttpFS supports Hadoop pseudo-authentication, HTTP SPNEGO Kerberos, and additional authentication mechanisms via a plugin API. HttpFS also supports Hadoop proxy user functionality.

The `webhdfs` client file system implementation can access HttpFS via the Hadoop filesystem command (`hadoop fs`), by using Hadoop DistCp, and from Java applications using the Hadoop file system Java API.

The HttpFS HTTP REST API is interoperable with the WebHDFS REST HTTP API.

For more information about HttpFS, see [Hadoop HDFS over HTTP](#).

The HttpFS role is required for Hue when you enable [HDFS high availability](#).

### Adding the HttpFS Role

1. Go to the HDFS service.
2. Click the **Instances** tab.
3. Click **Add Role Instances**.
4. Click the text box below the **HttpFS** field. The Select Hosts dialog displays.
5. Select the host on which to run the role and click **OK**.
6. Click **Continue**.
7. Check the checkbox next to the **HttpFS** role and select **Actions for Selected > Start**.

## Configuring CDH and Managed Services

### Using Load Balancer with HttpFS

To configure a load balancer, select **Clusters > Hive 2 > Configuration > Category > HttpFS** and enter the hostname and port number of the load balancer in the **HTTPFS Load Balancer** property in the format *hostname:port number*.

■ **Note:**

When you set this property, Cloudera Manager regenerates the keytabs for HttpFS roles. The principal in these keytabs contains the load balancer hostname.

If there is a Hue service that depends on this HDFS service, the Hue service has the option to use the load balancer as its HDFS Web Interface Role.

### Adding and Configuring an NFS Gateway

**Required Role:** **Cluster Administrator** **Full Administrator**

The NFS Gateway role implements an NFSv3 gateway. It is an optional role for a CDH 5 HDFS service.

The NFSv3 gateway allows a client to mount HDFS as part of the client's local file system. The gateway machine can be any host in the cluster, including the NameNode, a DataNode, or any HDFS client. The client can be any NFSv3-client-compatible machine.

After mounting HDFS to his or her local filesystem, a user can:

- Browse the HDFS file system as though it were part of the local file system
- Upload and download files from the HDFS file system to and from the local file system.
- Stream data directly to HDFS through the mount point.

File append is supported, but random write is not.

### Requirements and Limitations

- The NFS gateway works only with the following operating systems and Cloudera Manager and CDH versions:
  - With Cloudera Manager 5.0.1 or later and CDH 5.0.1 or later, the NFS gateway works on all operating systems supported by Cloudera Manager.
  - With Cloudera Manager 5.0.0 or CDH 5.0.0, the NFS gateway only works on RHEL and similar systems.
  - The NFS gateway is not supported on versions earlier than Cloudera Manager 5.0.0 and CDH 5.0.0.
- If any NFS server is already running on the NFS Gateway host, it must be stopped before the NFS Gateway role is started.
- There are two configuration options related to NFS Gateway role: **Temporary Dump Directory** and **Allowed Hosts and Privileges**. The **Temporary Dump Directory** is automatically created by the NFS Gateway role and should be configured before starting the role.
- The **Access Time Precision** property in the HDFS service must be enabled.

### Adding and Configuring the NFS Gateway Role

1. Go to the HDFS service.
2. Click the **Instances** tab.
3. Click **Add Role Instances**.
4. Click the text box below the **NFS Gateway** field. The Select Hosts dialog displays.
5. Select the host on which to run the role and click **OK**.
6. Click **Continue**.
7. Click the **NFS Gateway** role.
8. Click the **Configuration** tab.
9. Ensure that the requirements on the directory set in the **Temporary Dump Directory** property are met.



10. Optionally edit **Allowed Hosts and Privileges**.
11. Click the **Instances** tab.
12. Check the checkbox next to the **NFS Gateway** role and select **Actions for Selected > Start**.

### Setting HDFS Quotas

**Required Role:** **Cluster Administrator** **Full Administrator**

1. From the HDFS service page, select the **File Browser** tab.
2. Browse the file system to find the directory for which you want to set quotas.
3. Click the directory name so that it appears in the gray panel above the listing of its contents and in the detail section to the right of the File Browser table.
4. Click the **Edit Quota** button for the directory. A **Manage Quota** pop-up displays, where you can set file count or disk space limits for the directory you have selected.
5. When you have set the limits you want, click **OK**.

#### About file count limits

- The file count quota is a limit on the number of file and directory names in the directory configured.
- A directory counts against its own quota, so a quota of 1 forces the directory to remain empty.
- File counts are based on the intended replication factor for the files; changing the replication factor for a file will credit or debit quotas.

#### About disk space limits

- The space quota is a hard limit on the number of bytes used by files in the tree rooted at the directory being configured.
- Each replica of a block counts against the quota.
- The disk space quota calculation takes replication into account, so it uses the replicated size of each file, not the user-facing size.
- The disk space quota calculation includes open files (files presently being written), as well as files already written.
- Block allocations for files being written will fail if the quota would not allow a full block to be written.

### The Hive Service

There are two Hive service roles:

- **Hive Metastore Server** - manages the metastore process when Hive is configured with a remote metastore. You are strongly encouraged to read [Configuring the Hive Metastore \(CDH 4\)](#) or [Configuring the Hive Metastore \(CDH 5\)](#).
- **HiveServer2** - supports a Thrift API tailored for JDBC and ODBC clients, Kerberos authentication, and multi-client concurrency. There is also a CLI for HiveServer2 named Beeline. Cloudera recommends that you deploy HiveServer2 whenever possible. You can use the original HiveServer, and run it concurrently with HiveServer2. However, Cloudera Manager does not manage HiveServer, so you must configure and manage it outside Cloudera Manager. See [HiveServer2 documentation \(CDH 4\)](#) or [HiveServer2 documentation \(CDH 5\)](#) for more information.

#### The Hive Metastore Server

Cloudera recommends using a remote Hive metastore, especially for CDH 4.2 or later. Since the remote metastore is recommended, Cloudera Manager treats the Hive Metastore Server as a required role for all Hive services. Here are a couple key reasons why the remote metastore setup is advantageous, especially in production settings:

- The Hive metastore database password and JDBC drivers don't need to be shared with every Hive client; only the Hive Metastore Server does. Sharing passwords with many hosts is a security concern.
- You can control activity on the Hive metastore database. To stop all activity on the database, just stop the Hive Metastore Server. This makes it easy to perform tasks such as backup and upgrade, which require all Hive activity to stop.

## Configuring CDH and Managed Services

Information about the initial configuration of a remote Hive metastore database with Cloudera Manager can be found at [Cloudera Manager and Managed Service Data Stores](#).

### Considerations When Upgrading CDH

Hive has undergone major version changes from CDH 4.0 to 4.1 and between CDH 4.1 and 4.2. (CDH 4.0 had Hive 0.8.0, CDH 4.1 used Hive 0.9.0, and CDH 4.2 or later has 0.10.0). This requires that you manually back up and upgrade the Hive metastore database when upgrading between major Hive versions.

You should follow the steps in the appropriate in the Cloudera Manager procedure for upgrading CDH to upgrade the metastore *before* you restart the Hive service. This applies whether you are upgrading to packages or parcels. The procedure for upgrading CDH using packages is at [Upgrading CDH 4 Using Packages](#). The procedure for upgrading with parcels is at [Upgrading CDH 4 Using Parcels](#).

### Considerations When Upgrading Cloudera Manager

Cloudera Manager 4.5 added support for Hive, which includes the Hive Metastore Server role type. This role manages the metastore process when Hive is configured with a remote metastore.

When upgrading from Cloudera Manager prior to 4.5, Cloudera Manager automatically creates new Hive service(s) to capture the previous implicit Hive dependency from Hue and Impala. Your previous services will continue to function without impact. If Hue was using a Hive metastore backed by a Derby database, then the newly created Hive Metastore Server will also use Derby. Since Derby does not allow concurrent connections, Hue will continue to work, but the new Hive Metastore Server will fail to run. The failure is harmless (because nothing uses this new Hive Metastore Server at this point) and intentional, to preserve the set of cluster functionality as it was before upgrade. Cloudera discourages the use of a Derby backed Hive metastore due to its limitations. You should consider switching to a different supported database.

Cloudera Manager provides a Hive configuration option to bypass the Hive Metastore Server. When this configuration is enabled, Hive clients, Hue, and Impala connect directly to the Hive metastore database. Prior to Cloudera Manager 4.5, Hue and Impala connected directly to the Hive metastore database, so the bypass mode is enabled by default when upgrading to Cloudera Manager 4.5 or later. This is to ensure the upgrade doesn't disrupt your existing setup. You should plan to disable the bypass mode, especially when using CDH 4.2 or later. Using the Hive Metastore Server is the recommended configuration and the WebHCat Server role requires the Hive Metastore Server to *not* be bypassed. To disable bypass mode, see [Disabling Bypass Mode](#) on page 58.

Cloudera Manager 4.5 or later also supports HiveServer2 with CDH 4.2. In CDH 4 HiveServer2 is not added by default, but can be added as a new role under the Hive service (see [Role Instances](#) on page 40). In CDH 5, HiveServer2 is a mandatory role.

### Disabling Bypass Mode

**Required Role:** Configurator Cluster Administrator Full Administrator

In bypass mode Hive clients directly access the metastore database instead of using the Hive Metastore Server for metastore information.

1. Go to the Hive service.
2. Click the **Configuration** tab.
3. Expand the **Service-Wide > Advanced** category.
4. Uncheck the **Bypass Hive Metastore Server** checkbox.
5. Click **Save Changes**.
6. Re-deploy Hive client configurations.
7. Restart Hive and any Hue or Impala services configured to use that Hive service.

### Using Hive Gateways

Because the Hive service does not have worker roles, another mechanism is needed to enable the automatic propagation of client configurations to the other hosts in your cluster. Gateway roles fulfill this function. Gateways

in fact aren't really roles and do not have state, but they act as indicators for where client configurations should be placed. Hive gateways are created by default when the Hive service is added.

### Using a Load Balancer with HiveServer2

To configure a load balancer:

1. Go to the Hive service.
2. Click the **Configuration** tab.
3. Select **Category > HiveServer2 Default Group**
4. Enter the hostname and port number of the load balancer in the **HiveServer2 Load Balancer** property in the format `hostname:port number`.

■ **Note:**

When you set this property, Cloudera Manager regenerates the keytabs for HiveServer2 roles. The principal in these keytabs contains the load balancer hostname.

If there is a Hue service that depends on this Hive service, it also uses the load balancer to communicate with Hive.

### Hive Table Statistics

**Required Role:** Cluster Administrator Full Administrator

If your cluster has Impala then you can use the Impala implementation to compute statistics. The Impala implementation to compute table statistics is available in CDH 5.0.0 or higher and in Impala version 1.2.2 or higher. The Impala implementation of `COMPUTE STATS` requires no setup steps and is preferred over the Hive implementation. See [Overview of Table Statistics](#). If you are running an older version of Impala, you can collect statistics on a Hive table by running the following command from a Beeline client connected to HiveServer2:

```
analyze table <table name> compute statistics;
analyze table <table name> compute statistics for columns <all columns of a table>;
```

### Configuring Hive to Store Statistics in MySQL

If you are deploying CDH 5.2 this procedure is no longer required because Hive has another implementation of the statistics calculator.

By default, Hive writes statistics to a Derby database backed by a file named `/var/lib/hive/TempStatsStore`. However, in production systems Cloudera recommends that you store statistics in a database. Hive table statistics are not supported for PostgreSQL or Oracle. To configure Hive to store statistics in MySQL:

1. Set up a MySQL server. For instructions on setting up MySQL, see [MySQL Database](#).

This database will be heavily loaded, so it should not be installed on the same host as anything critical such as the Hive Metastore Server, the database hosting the Hive Metastore, or Cloudera Manager Server. When collecting statistics on a large table and/or in a large cluster, this host may become slow or unresponsive.

2. Create a statistics database in MySQL:

```
mysql> create database stats_db_name DEFAULT CHARACTER SET utf8;
Query OK, 1 row affected (0.00 sec)

mysql> grant all on stats_db_name.* TO 'stats_user'@'%' IDENTIFIED BY
'stats_password';
Query OK, 0 rows affected (0.00 sec)
```

3. Add the following into the **HiveServer2 Configuration Advanced Configuration Snippet for hive-site.xml** property:

- **Important:** Enter the contents of the `<value>` element as a single line without line breaks.

```
<property>
  <name>hive.stats.dbclass</name>
  <value>jdbc:mysql</value>
</property>
<property>
  <name>hive.stats.jdbcdriver</name>
  <value>com.mysql.jdbc.Driver</value>
</property>
<property>
  <name>hive.stats.dbconnectionstring</name>
  <value>jdbc:mysql://<stats_mysql_host>:3306/<stats_db_name>
    ?useUnicode=true&characterEncoding=UTF-8&
    user=<stats_user>&password=<stats_password></value>
</property>
<property>
  <name>hive.aux.jars.path</name>
  <value>file:///usr/share/java/mysql-connector-java.jar</value>
</property>
```

4. Click **Save Changes** to commit the changes.
5. Restart the HiveServer2 role.

## Using User-Defined Functions (UDFs) with HiveServer2

The `ADD JAR` command does *not* work with HiveServer2 and the Beeline client when Beeline runs on a different host. As an alternative to `ADD JAR`, Hive's *auxiliary paths* functionality should be used. Perform one of the following procedures depending on whether you want to create permanent or temporary functions.

### User-Defined Functions (UDFs) with HiveServer2 Using Cloudera Manager

**Required Role:** Configurator Cluster Administrator Full Administrator

#### Creating Permanent Functions

1. Copy the JAR file to HDFS and make sure the `hive` user can access this JAR file.
2. Copy the JAR file to the host on which HiveServer2 is running. Save the JARs to any directory you choose, give the `hive` user read, write, and execute access to this directory, and make a note of the path (for example, `/opt/local/hive/lib/`).
3. In the Cloudera Manager Admin Console, go to the Hive service.
4. Click the **Configuration** tab.
5. Expand the **Service-Wide > Advanced** categories.
6. Configure the **Hive Auxiliary JARs Directory** property with the HiveServer2 host path from Step 1, `/opt/local/hive/lib/`.
7. Click **Save Changes**. The JARs are added to `HIVE_AUX_JARS_PATH` environment variable.
8. Redeploy the Hive client configuration.
  - a. In the Cloudera Manager Admin Console, go to the Hive service.
  - b. From the **Actions** menu at the top right of the service page, select **Deploy Client Configuration**.
  - c. Click **Deploy Client Configuration**.
9. Restart the Hive service. If the **Hive Auxiliary JARs Directory** property is configured but the directory does not exist, HiveServer2 will not start.
10. **If Sentry is enabled** - Grant privileges on the JAR files to the roles that require access. Login to Beeline as user `hive` and use the Hive SQL [GRANT](#) statement to do so. For example:

```
GRANT ALL ON URI 'file:///opt/local/hive/lib/my.jar' TO ROLE EXAMPLE_ROLE
```

You must also grant privilege to the JAR on HDFS:

```
GRANT ALL ON URI 'hdfs:///path/to/jar' TO ROLE EXAMPLE_ROLE
```

11. Run the `CREATE FUNCTION` command and point to the JAR file location in HDFS. For example:

```
CREATE FUNCTION addfunc AS 'com.example.hiveserver2.udf.add' USING JAR
'hdfs:///path/to/jar'
```

### Creating Temporary Functions

1. Copy the JAR file to the host on which HiveServer2 is running. Save the JARs to any directory you choose, give the `hive` user read, write, and execute access to this directory, and make a note of the path (for example, `/opt/local/hive/lib/`).
2. In the Cloudera Manager Admin Console, go to the Hive service.
3. Click the **Configuration** tab.
4. Expand the **Service-Wide > Advanced** categories.
5. Configure the **Hive Auxiliary JARs Directory** property with the HiveServer2 host path from Step 1, `/opt/local/hive/lib/`.
6. Click **Save Changes**. The JARs are added to `HIVE_AUX_JARS_PATH` environment variable.
7. Redeploy the Hive client configuration.
  - a. In the Cloudera Manager Admin Console, go to the Hive service.
  - b. From the **Actions** menu at the top right of the service page, select **Deploy Client Configuration**.
  - c. Click **Deploy Client Configuration**.
8. Restart the Hive service. If the **Hive Auxiliary JARs Directory** property is configured but the directory does not exist, HiveServer2 will not start.
9. **If Sentry is enabled** - Grant privileges on the local JAR files to the roles that require access. Login to Beeline as user `hive` and use the Hive SQL [GRANT](#) statement to do so. For example:

```
GRANT ALL ON URI 'file:///opt/local/hive/lib/my.jar' TO ROLE EXAMPLE_ROLE
```

10. Run the `CREATE TEMPORARY FUNCTION` command. For example:

```
CREATE TEMPORARY FUNCTION addfunc AS 'com.example.hiveserver2.udf.add'
```

### User-Defined Functions (UDFs) with HiveServer2 Using the Command Line

The following sections describe how to create permanent and temporary functions using the command line.

#### Creating Permanent Functions

1. Copy the JAR file to HDFS and make sure the `hive` user can access this JAR file.
2. On the Beeline client machine, in `/etc/hive/conf/hive-site.xml`, set the `hive.aux.jars.path` property to a comma-separated list of the fully-qualified paths to the JAR file and any dependent libraries.

```
hive.aux.jars.path=file:///opt/local/hive/lib/my.jar
```

3. Copy the JAR file (and its dependent libraries) to the host running HiveServer2/Impala. Make sure the `hive` user has read, write, and execute access to these files on the HiveServer2/Impala host.
4. On the HiveServer2/Impala host, open `/etc/default/hive-server2` and set the `AUX_CLASSPATH` variable to a comma-separated list of the fully-qualified paths to the JAR file and any dependent libraries.

```
AUX_CLASSPATH=/opt/local/hive/lib/my.jar
```

5. Restart HiveServer2.

## Configuring CDH and Managed Services

6. **If Sentry is enabled** - Grant privileges on the JAR files to the roles that require access. Login to Beeline as user `hive` and use the Hive SQL [GRANT](#) statement to do so. For example:

```
GRANT ALL ON URI 'file:///opt/local/hive/lib/my.jar' TO ROLE EXAMPLE_ROLE
```

You must also grant privilege to the JAR on HDFS:

```
GRANT ALL ON URI 'hdfs:///path/to/jar' TO ROLE EXAMPLE_ROLE
```

If you are using Sentry policy files, you can grant the URI privilege as follows:

```
udf_r = server=server1->uri=file:///opt/local/hive/lib  
udf_r = server=server1->uri=hdfs:///path/to/jar
```

7. Run the `CREATE FUNCTION` command and point to the JAR from Hive:

```
CREATE FUNCTION addfunc AS 'com.example.hiveserver2.udf.add' USING JAR  
'hdfs:///path/to/jar'
```

### Creating Temporary Functions

1. On the Beeline client machine, in `/etc/hive/conf/hive-site.xml`, set the `hive.aux.jars.path` property to a comma-separated list of the fully-qualified paths to the JAR file and any dependent libraries.

```
hive.aux.jars.path=file:///opt/local/hive/lib/my.jar
```

2. Copy the JAR file (and its dependent libraries) to the host running HiveServer2/Impala. Make sure the `hive` user has read, write, and execute access to these files on the HiveServer2/Impala host.
3. On the HiveServer2/Impala host, open `/etc/default/hive-server2` and set the `AUX_CLASSPATH` variable to a comma-separated list of the fully-qualified paths to the JAR file and any dependent libraries.

```
AUX_CLASSPATH=/opt/local/hive/lib/my.jar
```

4. **If Sentry is enabled** - Grant privileges on the local JAR files to the roles that require access. Login to Beeline as user `hive` and use the Hive SQL [GRANT](#) statement to do so. For example:

```
GRANT ALL ON URI 'file:///opt/local/hive/lib/my.jar' TO ROLE EXAMPLE_ROLE
```

If you are using Sentry policy files, you can grant the URI privilege as follows:

```
udf_r = server=server1->uri=file:///opt/local/hive/lib
```

5. Restart HiveServer2.
6. Run the `CREATE FUNCTION` command and point to the JAR from Hive:

```
CREATE FUNCTION addfunc AS 'com.example.hiveserver2.udf.add'
```

## The Hue Service

Hue is a set of web UIs that enable you to interact with a CDH cluster.

### Configuring Hue to Work with High Availability

If your cluster has high availability enabled, you must configure the Hue HDFS Web Interface Role property to use HTTPFS. See [Configuring Hue to Work with HDFS HA](#) on page 233 for detailed instructions.

### Managing Hue Analytics Data Collection

Required Role: **Configurator** **Cluster Administrator** **Full Administrator**

Hue tracks anonymised pages and application versions in order to gather information to help compare each application's usage levels. The data collected does not include any hostnames or IDs. For example, the data is of the form: /2.3.0/pig, /2.5.0/beeswax/execute. You can restrict data collection as follows:

1. Go to the Hue service.
2. Click the **Configuration** tab.
3. Expand the **Service-Wide** category.
4. Uncheck the **Enable Usage Data Collection** checkbox.
5. Click **Save Changes** to commit the changes.
6. Restart the Hue service.

### Using an External Database for Hue

Required Role: **Full Administrator**

By default, Cloudera Manager uses SQLite for the Hue database. If necessary, you can configure Cloudera Manager to use an external database such as MySQL, PostgreSQL, or Oracle as the database for Hue. The databases that Hue supports are listed at:

- [CDH 4 supported databases](#)
- [CDH 5 supported databases](#)

- **Note:** In the instructions that follow, dumping the database and editing the JSON objects is only necessary if you have data in SQLite that you need to migrate. If you don't need to migrate data from SQLite, you can skip those steps.

### Configuring the Hue Server to Store Data in MySQL

- **Note:** Cloudera recommends you use InnoDB, *not* MyISAM, as your MySQL engine.

1. In the Cloudera Manager Admin Console, go to the Hue service status page.
2. Select **Actions** > **Stop**. Confirm you want to stop the service by clicking **Stop**.
3. In the Status Summary, click a Hue Server instance.
4. Select **Actions** > **Dump Database**. Confirm you want to dump the database by clicking **Dump Database**.
5. Open the database dump file (by default /tmp/hue\_database\_dump.json) and remove all JSON objects with `useradmin.userprofile` in the `model` field. (You can verify the location of the database dump file by searching for Database Dump File in the Hue configuration settings.)
6. Configure MySQL to set strict mode:

```
[mysqld]
sql_mode=STRICT_ALL_TABLES
```

7. Create a new database and grant privileges to a Hue user to manage this database. For example:

```
mysql> create database hue CHARACTER SET utf8 COLLATE utf8_general_cs;
Query OK, 1 row affected (0.01 sec)
mysql> grant all on hue.* to 'hue'@'localhost' identified by 'secretpassword';
Query OK, 0 rows affected (0.00 sec)
```

8. In the Cloudera Manager Admin Console, click the Hue service.
9. Click the **Configuration** tab.
10. Expand the **Service-Wide** > **Database** category.
11. Specify the settings for **Hue Database Type**, **Hue Database Hostname**, **Hue Database Port**, **Hue Database Username**, **Hue Database Password**, and **Hue Database Name**. For example, for a MySQL database on the local host, you might use the following values:
  - Hue Database Type = `mysql`
  - Hue Database Hostname = `host`

## Configuring CDH and Managed Services

- Hue Database Port = 3306
- Hue Database Username = hue
- Hue Database Password = *secretpassword*
- Hue Database Name = hue

12. Optionally restore the Hue data to the new database:

a. Select **Actions** > **Synchronize Database**.

b. Determine the foreign key ID.

```
$ mysql -uhue -psecretpassword
mysql > SHOW CREATE TABLE auth_permission;
```

c. (InnoDB only) Drop the foreign key that you retrieved in the previous step.

```
mysql > ALTER TABLE auth_permission DROP FOREIGN KEY
content_type_id_refs_id_XXXXXX;
```

d. Delete the rows in the `django_content_type` table.

```
mysql > DELETE FROM hue.django_content_type;
```

e. In Hue service instance page, click **Actions** > **Load Database**. Confirm you want to load the database by clicking **Load Database**.

f. (InnoDB only) Add back the foreign key.

```
mysql > ALTER TABLE auth_permission ADD FOREIGN KEY (content_type_id) REFERENCES
django_content_type (id);
```

13. Start the Hue service.

### Configuring the Hue Server to Store Data in PostgreSQL

1. In the Cloudera Manager Admin Console, go to the Hue service status page.
2. Select **Actions** > **Stop**. Confirm you want to stop the service by clicking **Stop**.
3. In the Status Summary, click a Hue Server instance.
4. Select **Actions** > **Dump Database**. Confirm you want to dump the database by clicking **Dump Database**.
5. Open the database dump file (by default `/tmp/hue_database_dump.json`) and remove all JSON objects with `useradmin.userprofile` in the `model` field. (You can verify the location of the database dump file by searching for Database Dump File in the Hue configuration settings.)
6. Install required packages.

#### RHEL

```
$ sudo yum install postgresql-devel gcc python-devel
```

#### SLES

```
$ sudo zypper install postgresql-devel gcc python-devel
```

#### Ubuntu or Debian

```
$ sudo apt-get install postgresql-devel gcc python-devel
```

7. Install the Python module that provides the connector to PostgreSQL:



- **Parcel install**

```
$ sudo /opt/cloudera/parcels/CDH/lib/hue/build/env/bin/pip install setuptools
$ sudo /opt/cloudera/parcels/CDH/lib/hue/build/env/bin/pip install psycopg2
```

- **Package install**

- **CDH 4**

```
sudo -u hue /usr/share/hue/build/env/bin/pip install setuptools
sudo -u hue /usr/share/hue/build/env/bin/pip install psycopg2
```

- **CDH 5**

```
sudo -u hue /usr/lib/hue/build/env/bin/pip install setuptools
sudo -u hue /usr/lib/hue/build/env/bin/pip install psycopg2
```

## 8. Install the PostgreSQL server.

### RHEL

```
$ sudo yum install postgresql-server
```

### SLES

```
$ sudo zypper install postgresql-server
```

### Ubuntu or Debian

```
$ sudo apt-get install postgresql
```

## 9. Initialize the data directories.

```
$ service postgresql initdb
```

## 10. Configure client authentication.

- Edit `/var/lib/pgsql/data/pg_hba.conf`.
- Set the authentication methods for local to `trust` and for host to `password` and add the following line at the end.

```
host hue hue 0.0.0.0/0 md5
```

## 11. Start the PostgreSQL server.

```
$ su - postgres
# /usr/bin/postgres -D /var/lib/pgsql/data > logfile 2>&1 &
```

## 12. Configure PostgreSQL to listen on all network interfaces.

- Edit `/var/lib/pgsql/data/postgresql.conf` and set `listen_addresses`.

```
listen_addresses = '0.0.0.0'      # Listen on all addresses
```

## 13. Create the hue database and grant privileges to a hue user to manage the database.

```
# psql -U postgres
postgres=# create database hue;
postgres=# \c hue;
```

```
You are now connected to database 'hue'.
postgres=# create user hue with password 'secretpassword';
postgres=# grant all privileges on database hue to hue;
postgres=# \q
```

14. Restart the PostgreSQL server.

```
$ sudo service postgresql restart
```

15. Verify connectivity.

```
psql -h localhost -U hue -d hue
Password for user hue: secretpassword
```

16. Configure the PostgreSQL server to start at boot.

### RHEL

```
$ sudo /sbin/chkconfig postgresql on
$ sudo /sbin/chkconfig --list postgresql
postgresql          0:off  1:off  2:on   3:on   4:on   5:on   6:off
```

### SLES

```
$ sudo chkconfig --add postgresql
```

### Ubuntu or Debian

```
$ sudo chkconfig postgresql on
```

17. In the Cloudera Manager Admin Console, click the Hue service.

18. Click the **Configuration** tab.

19. In the **Category** pane, click **Advanced** under **Service-Wide**.

20. Specify the settings for **Hue Server Configuration Advanced Configuration Snippet**:

```
[desktop]
[[database]]
host=localhost
port=5432
engine=postgresql_psycopg2
user=hue
password=secretpassword
name=hue
```

- **Note:** If you specify the database host, port, username, password, and name in the respective **Service-Wide > Database > Hue Database \*** properties, you can omit those properties from the above configuration. In particular, you can avoid storing the password in the Hue configuration file in plain text.

21. Click **Save Changes**.

22. Optionally restore the Hue data to the new database:

- a. Select **Actions > Synchronize Database**.

- b. Determine the foreign key ID.

```
bash# su - postgres
$ psql -h localhost -U hue -d hue
postgres=# \d auth_permission;
```

- c. Drop the foreign key that you retrieved in the previous step.

```
postgres=# ALTER TABLE auth_permission DROP CONSTRAINT
content_type_id_refs_id_XXXXXX;
```

- d. Delete the rows in the `django_content_type` table.

```
postgres=# TRUNCATE django_content_type CASCADE;
```

- e. In Hue service instance page, **Actions** > **Load Database**. Confirm you want to load the database by clicking **Load Database**.

- f. Add back the foreign key you dropped.

```
bash# su - postgres
$ psql -h localhost -U hue -d hue
postgres=# ALTER TABLE auth_permission ADD CONSTRAINT
content_type_id_refs_id_XXXXXX FOREIGN KEY (content_type_id) REFERENCES
django_content_type(id) DEFERRABLE INITIALLY DEFERRED;
```

23. Start the Hue service.

### *Configuring the Hue Server to Store Data in Oracle (Parcel Installation)*

Use the following instructions to configure the Hue Server with an Oracle database if you are working on a parcel-based deployment. If you are using packages, see [Configuring the Hue Server to Store Data in Oracle \(Package Installation\)](#) on page 68.

1. Ensure Python 2.6 or newer is installed on the server Hue is running on.
2. Add <http://tiny.cloudera.com/hue-oracle-client-db> to the Cloudera Manager remote parcel repository URL list and download, distribute, and activate the parcel.
3. For CDH versions lower than 5.3, install the Python Oracle library: `$ HUE_HOME/build/env/bin/pip install cx_Oracle`
4. Upgrade django south.

```
$ <HUE_HOME>/build/env/bin/pip install south --upgrade
```

5. In the Cloudera Manager Admin Console, go to the Hue service status page.
6. Select **Actions** > **Stop**. Confirm you want to stop the service by clicking **Stop**.
7. In the Status Summary, click a Hue Server instance.
8. Select **Actions** > **Dump Database**. Confirm you want to dump the database by clicking **Dump Database**.
9. Click the **Configuration** tab.
10. Set the **Hue Service Advanced Configuration Snippet (Safety Valve)** for `hue_safety_valve.ini` property. Add the following options (and modify accordingly for your setup):

```
[desktop]
[[database]]
host=localhost
port=1521
engine=oracle
user=hue
password=secretpassword
name=<SID of the Oracle database, for example, 'XE'>
```

For CDH 5.1 and higher you can use an Oracle service name. To use the Oracle service name instead of the SID, use the following configuration instead:

```
port=0
engine=oracle
user=hue
```

```
password=secretpassword
name=oracle.example.com:1521/orcl.example.com
```

The directive `port=0` allows Hue to use a service name. The `name` string is the connect string, including hostname, port, and service name.

- **Note:** If you specify the database host, port, username, password, and name in the respective **Service-Wide > Database > Hue Database \*** properties, you can omit those properties from the above configuration. In particular, you can avoid storing the password in the Hue configuration file in plain text.

To add support for a multithreaded environment, set the `threaded` option to `true` under the `[desktop]>[[database]]` section.

```
options={'threaded':true}
```

### 11. Grant required permissions to the Hue user in Oracle:

```
grant alter any index to hue;
grant alter any table to hue;
grant alter database link to hue;
grant create any index to hue;
grant create any sequence to hue;
grant create database link to hue;
grant create session to hue;
grant create table to hue;
grant drop any sequence to hue;
grant select any dictionary to hue;
grant drop any table to hue;
grant create procedure to hue;
grant create trigger to hue;
```

### 12. Navigate to the Hue Server instance in Cloudera Manager and select **Actions > Synchronize Database**.

### 13. Generate statements to delete all data from Oracle tables:

- If you are connected to Oracle as the Hue user, run the following command:

```
SELECT 'DELETE FROM ' || table_name || ';' FROM user_tables;
```

- If you are connected to Oracle as a user that is not the Hue user, run the following command:

```
SELECT 'DELETE FROM ' || '<schema_name>.' || table_name || ';' FROM user_tables;
```

Where `<schema_name>` is the correct hue schema, typically: `hue`.

### 14. Run the statements generated in the preceding step.

### 15. Load the data that you dumped. Navigate to the Hue Server instance and select **Actions > Load Database**. This step is not necessary if you have a fresh Hue install with no data or if you don't want to save the Hue data.

### 16. Start the Hue service.

## Configuring the Hue Server to Store Data in Oracle (Package Installation)

If you have a parcel-based environment, see [Configuring the Hue Server to Store Data in Oracle \(Parcel Installation\)](#) on page 67.

1. Download the Oracle libraries at [Instant Client for Linux x86-64 Version 11.1.0.7.0](#), Basic and SDK (with headers) zip files to the same directory.
2. Unzip the Oracle client zip files.

- Set environment variables to reference the libraries.

```
$ export ORACLE_HOME=oracle_download_directory
$ export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:$ORACLE_HOME
```

- Create a symbolic link for the shared object:

```
$ cd $ORACLE_HOME
$ ln -sf libclntsh.so.11.1 libclntsh.so
```

- Ensure Python 2.6 or newer is installed on the server Hue is running on.
- For CDH versions lower than 5.3, install the Python Oracle library: `$ HUE_HOME/build/env/bin/pip install cx_Oracle`
- Upgrade django south: `$ HUE_HOME/build/env/bin/pip install south --upgrade`
- In the Cludera Manager Admin Console, go to the Hue service status page.
- Select **Actions** > **Stop**. Confirm you want to stop the service by clicking **Stop**.
- In the Status Summary, click a Hue Server instance.
- Select **Actions** > **Dump Database**. Confirm you want to dump the database by clicking **Dump Database**.
- Click the **Configuration** tab.
- Expand the **Service-Wide** > **Advanced** category.
- Set the **Hue Service Environment Advanced Configuration Snippet (Safety Valve)** property to

```
ORACLE_HOME=oracle_download_directory
LD_LIBRARY_PATH=$LD_LIBRARY_PATH:oracle_download_directory
```

- Set the **Hue Service Advanced Configuration Snippet (Safety Valve)** for `hue_safety_valve.ini` property. Add the following options (and modify accordingly for your setup):

```
[desktop]
[[database]]
host=localhost
port=1521
engine=oracle
user=hue
password=secretpassword
name=<SID of the Oracle database, for example, 'XE'>
```

For CDH 5.1 and higher you can use an Oracle service name. To use the Oracle service name instead of the SID, use the following configuration instead:

```
port=0
engine=oracle
user=hue
password=secretpassword
name=oracle.example.com:1521/orcl.example.com
```

The directive `port=0` allows Hue to use a service name. The `name` string is the connect string, including hostname, port, and service name.

- Note:** If you specify the database host, port, username, password, and name in the respective **Service-Wide** > **Database** > **Hue Database \*** properties, you can omit those properties from the above configuration. In particular, you can avoid storing the password in the Hue configuration file in plain text.

To add support for a multithreaded environment, set the `threaded` option to `true` under the `[desktop]>[[database]]` section.

```
options={'threaded':true}
```

16. Grant required permissions to the Hue user in Oracle:

```
grant alter any index to hue;
grant alter any table to hue;
grant alter database link to hue;
grant create any index to hue;
grant create any sequence to hue;
grant create database link to hue;
grant create session to hue;
grant create table to hue;
grant drop any sequence to hue;
grant select any dictionary to hue;
grant drop any table to hue;
grant create procedure to hue;
grant create trigger to hue;
```

17. Navigate to the Hue Server instance in Cloudera Manager and select **Actions > Synchronize Database**.

18. Generate statements to delete all data from Oracle tables:

- If you are connected to Oracle as the Hue user, run the following command:

```
SELECT 'DELETE FROM ' || table_name || ';' FROM user_tables;
```

- If you are connected to Oracle as a user that is not the Hue user, run the following command:

```
SELECT 'DELETE FROM ' || '<schema_name>.' || table_name || ';' FROM user_tables;
```

Where *<schema\_name>* is the correct hue schema, typically: hue.

19. Run the statements generated in the preceding step.

20. Load the data that you dumped. Navigate to the Hue Server instance and select **Actions > Load Database**. This step is not necessary if you have a fresh Hue install with no data or if you don't want to save the Hue data.

21. Start the Hue service.

### Enabling Hue Applications

**Required Role:** Configurator Cluster Administrator Full Administrator

Most Hue applications are configured by default, based on the services you have installed. Cloudera Manager selects the service instance that Hue depends on. If you have more than one service, you may want to verify or change the service dependency for Hue. Also, if you add a service such as Sqoop 2 or Oozie after you have set up Hue, you will need to set the dependency because it won't be done automatically. To add a dependency:

1. Go to the **Hue** service.
2. Click the **Configuration** tab.
3. Select the **Service-Wide** category.
4. Change the setting for the service dependency from **None** to the appropriate service instance.
5. Click **Save Changes**.
6. Restart the Hue service.

### Enabling the Sqoop 2 Application

If you upgrade to Cloudera Manager 4.7 from an earlier version of Cloudera Manager 4, you will need to set the Hue dependency to enable the Sqoop 2 application.

### Enabling the HBase Browser Application

**Required Role:** Full Administrator

The HBase Browser application, new as of CDH 4.4, depends on the HBase Thrift server for its functionality.

1. [Add the HBase Thrift Server role.](#)
2. Configure Hue to point to the Thrift Server:
  - a. Select the **Hue** service.
  - b. Click the **Configuration** tab.
  - c. Go to the **Service-Wide** category.
  - d. For the **HBase Service** property, make sure it is set to the HBase service for which you enabled the Thrift Server role (if you have more than one HBase service instance).
  - e. In the **HBase Thrift Server** property, click in the edit field and select the Thrift Server role that Hue should use.
  - f. **Save Changes.**

### Enabling the Solr Search Application

To use the Solr Search application with Hue, you must update the URL for the Solr Server in the Hue Server advanced configuration snippet. In addition, if you are using parcels with CDH 4.3, you must register the "hue-search" application manually or access will fail. You do not need to do this if you are using CDH 4.4 or later. See [Deploying Solr with Hue](#) on page 92 for detailed instructions.

### The Impala Service

You can install Cloudera Impala through the Cloudera Manager installation wizard, using either parcels or packages, and have the service created and started as part of the Installation wizard. See [Installing Impala](#).

If you elect not to include the Impala service using the Installation wizard, you can use the **Add Service** wizard to perform the installation. The wizard will automatically configure and start the dependent services and the Impala service. See [Adding a Service](#) on page 30 for instructions.

For further information on the Impala service, see:

- **CDH 4** - [Cloudera Impala Documentation for CDH 4](#)
- **CDH 5** - [Cloudera Impala Guide](#)

For information on features that support Impala resource management see [Impala Resource Management](#) on page 188.

### Configuring the Impala Service

There are several types of configuration settings you may need to apply, depending on your situation.

#### Running Impala with CDH 4.1

If you are running CDH 4.1, and the Bypass Hive Metastore Server option is enabled, do the following:

1. Go to the Impala service.
2. Click the **Configuration** tab.
3. Select **Impala Daemon Default Group > Advanced**.
4. Add the following to the **Impala Advanced Configuration Snippet for hive-site.xml** property, replacing *hive\_metastore\_server\_host* with the name of your Hive Metastore Server host:

```
<property>
  <name>hive.metastore.local</name>
  <value>false</value>
</property>
<property>
  <name>hive.metastore.uris</name>
  <value>thrift://hive_metastore_server_host:9083</value>
</property>
```

5. Click **Save Changes**.

## Configuring CDH and Managed Services

6. Restart the Impala service.

### Configuring Table Statistics

Configuring table statistics is highly recommended when using Impala. It allows Impala to make optimizations that can result in significant (over 10x) performance improvement for some joins. If these are not available, Impala will still function, but at lower performance.

The Impala implementation to compute table statistics is available in CDH 5.0.0 or higher and in Impala version 1.2.2 or higher. The Impala implementation of `COMPUTE STATS` requires no setup steps and is preferred over the Hive implementation. See [Overview of Table Statistics](#). If you are running an older version of Impala, follow the procedure in [Hive Table Statistics](#) on page 59.

### Using a Load Balancer with Impala

To configure a load balancer:

1. Go to the Impala service.
2. Click the **Configuration** tab.
3. Select **Category > Impala Daemon Default Group**
4. Enter the hostname and port number of the load balancer in the **Impala Daemons Load Balancer** property in the format `hostname:port number`.

■ **Note:**

When you set this property, Cloudera Manager regenerates the keytabs for Impala Daemon roles. The principal in these keytabs contains the load balancer hostname.

If there is a Hue service that depends on this Impala service, it also uses the load balancer to communicate with Impala.

### Impala Web Servers

#### Enabling and Disabling Access to Impala Web Servers

Each of the Impala-related daemons includes a built-in web server that lets an administrator diagnose issues with each daemon on a particular host, or perform other administrative actions such as cancelling a running query. By default, these web servers are enabled. You might turn them off in a high-security configuration where it is not appropriate for users to have access to this kind of monitoring information through a web interface. (To leave the web servers enabled but control who can access their web pages, consult the *Configuring Secure Access for Impala Web Servers* later in this section.)

- **Impala StateStore**
  1. Go to the Impala service.
  2. Click the **Configuration** tab.
  3. Select **Impala StateStore Default Group**.
  4. Check or uncheck **Enable StateStore Web Server**.
  5. Click **Save Changes**.
  6. Restart the Impala service.
- **Impala Daemon**
  1. Go to the Impala service.
  2. Click the **Configuration** tab.
  3. Select **Impala Daemon Default Group > Ports and Addresses**.
  4. Check or uncheck **Enable Impala Daemon Web Server**.
  5. Click **Save Changes**.
  6. Restart the Impala service.



### Opening Impala Web Server UIs

- **Impala StateStore**
  1. Go to the Impala service.
  2. Select **Web UI > Impala StateStore Web UI**.
- **Impala Daemon**
  1. Go the to Impala service.
  2. Click the **Instances** tab.
  3. Click an **Impala Daemon** instance.
  4. Click **Impala Daemon Web UI**.
- **Impala Catalog Server**
  1. Go to the Impala service.
  2. Select **Web UI > Impala Catalog Web UI**.
- **Impala Llama ApplicationMaster**
  1. Go to the Impala service.
  2. Click the **Instances** tab.
  3. Click a **Impala Llama ApplicationMaster** instance.
  4. Click **Llama Web UI**.

### Configuring Secure Access for Impala Web Servers

Cloudera Manager supports two methods of authentication for secure access to the Impala Catalog Server, Daemon, and StateStoreweb servers: password-based authentication and SSL certificate authentication. Both of these can be configured through properties of the Impala Catalog Server, Daemon, and StateStore. Authentication for the three types of daemons can be configured independently.

#### Configuring Password Authentication

1. Go to the Impala service.
2. Click the **Configuration** tab.
3. Search for "password" using the Search box within the Configuration page. This should display the password-related properties (Username and Password properties) for the Impala Catalog Server, Daemon, and StateStore. If there are multiple role groups configured for Impala Daemon instances, the search should display all of them.
4. Enter a username and password into these fields.
5. Click **Save Changes**.
6. Restart the Impala service.

Now when you access the Web UI for the Impala Catalog Server, Daemon, and StateStore, you are asked to log in before access is granted.

#### Configuring SSL Certificate Authentication

1. Create or obtain an SSL certificate.
2. Place the certificate, in `.pem` format, on the hosts where the Impala Catalog Server and StateStore are running, and on each host where an Impala Daemon is running. It can be placed in any location (path) you choose. If all the Impala Daemons are members of the same role group, then the `.pem` file must have the same path on every host.
3. Go to the Impala service page.
4. Click the **Configuration** tab.

## Configuring CDH and Managed Services

5. Search for "certificate" using the Search box within the Configuration page. This should display the certificate file location properties for the Impala Catalog Server, Daemon, and StateStore. If there are multiple role groups configured for Impala Daemon instances, the search should display all of them.
6. In the property fields, enter the full path name to the certificate file.
7. Click **Save Changes**.
8. Restart the Impala service.

When you access the Web UI for the Impala Catalog Server, Daemon, and StateStore, `https` will be used.

### The Isilon Service

EMC Isilon is a storage service with a distributed file system that can be used in place of HDFS to provide storage for CDH services.

- **Note:** This documentation covers only the Cloudera Manager portion of using EMC Isilon storage with Cloudera Manager. For information about tasks performed on Isilon OneFS, see the information hub for Cloudera on the EMC Community Network: <https://community.emc.com/docs/DOC-39522>.

### Supported Versions

The following versions of Cloudera and Isilon products are supported:

- Cloudera Manager 5.2
- CDH 5.1.3
- Isilon OneFS 7.2.0

- **Note:** The Impala service and Cloudera Navigator are not supported in this release.

### Preliminary Steps on the Isilon Service

Before installing a Cloudera Manager cluster to use Isilon storage, perform the following steps on the Isilon OneFS system. For detailed information on setting up Isilon OneFS for Cloudera Manager, see the Isilon documentation at <https://community.emc.com/docs/DOC-39522>.

1. Create an Isilon access zone with HDFS support.

```
Example:
/ifs/your-access-zone/hdfs
```

- **Note:** The above is simply an example; the HDFS root directory does not have to begin with `ifs` or end with `hdfs`.

2. Create two directories that will be used by all CDH services:

- a. Create a `tmp` directory in the access zone.

- Create `supergroup` group and `hdfs` user.
- Create a `tmp` directory and set ownership to `hdfs:supergroup`, and permissions to `1777`.

```
Example:
cd hdfs_root_directory
isi_run -z zone_id mkdir tmp
isi_run -z zone_id chown hdfs:supergroup tmp
isi_run -z zone_id chmod 1777 tmp
```

- b. Create a `user` directory in the access zone and set ownership to `hdfs:supergroup`, and permissions to `755`

```
Example:
cd hdfs_root_directory
```

```
isi_run -z zone_id mkdir user
isi_run -z zone_id chown hdfs:supergroup user
isi_run -z zone_id chmod 755 user
```

3. Create the service-specific users, groups, or directories for each CDH service you plan to use. Create the directories under the access zone you have created.

- **Note:** Many of the values provided in the examples below are default values in Cloudera Manager and must match the Cloudera Manager configuration settings. The format for the examples is: *dir user: group permission* . Create the directories below under the access zone you have created, for example, */ifs/ your-access-zone /hdfs/*

- ZooKeeper: nothing required.
- HBase
  - Create hbase group with hbase user.
  - Create the root directory for HBase:

```
Example:
hdfs_root_directory/hbase hbase:hbase 755
```

- YARN (MR2)
  - Create mapred group with mapred user.
  - Create history directory for YARN:

```
Example:
hdfs_root_directory/user/history mapred:hadoop 777
```

- Create the remote application log directory for YARN:

```
Example:
hdfs_root_directory/tmp/logs mapred:hadoop 775
```

- Oozie
  - Create oozie group with oozie user.
  - Create the user directory for Oozie:

```
Example:
hdfs_root_directory/user/oozie oozie:oozie 775
```

- Hive
  - Create hive group with hive user.
  - Create the user directory for Hive:

```
Example:
hdfs_root_directory/user/hive hive:hive 775
```

- Create the warehouse directory for Hive:

```
Example:
hdfs_root_directory/user/hive/warehouse hive:hive 1777
```

## Configuring CDH and Managed Services

- Solr
  - Create `solr` group with `solr` user.
  - Create the data directory for Solr:

```
Example:  
hdfs_root_directory/solr solr:solr 775
```

- Sqoop
  - Create `sqoop` group with `sqoop2` user.
  - Create the user directory for Sqoop:

```
Example:  
hdfs_root_directory/user/sqoop2 sqoop2:sqoop 775
```

- Hue
  - Create `hue` group with `hue` user.
  - Create `sample` group with `sample` user.

- Spark
  - Create `spark` group with `spark` user.
  - Create the user directory for Spark:

```
Example:  
hdfs_root_directory/user/spark spark:spark 751
```

- Create application history directory for Spark:

```
Example:  
hdfs_root_directory/user/spark/applicationHistory spark:spark 1777
```

Once the users, groups, and directories are created in Isilon OneFS, you are ready to install Cloudera Manager.

### Installing Cloudera Manager with Isilon

To install Cloudera Manager following the instructions provided in [Cloudera Installation and Upgrade](#).

- The simplest installation procedure, suitable for development or proof of concept, is [Installation Path A](#), which uses embedded databases that are installed as part of the Cloudera Manager installation process.
- For production environments, [Installation Path B](#) describes configuring external databases for Cloudera Manager and CDH storage needs.

If you choose parcel installation on the **Cluster Installation** screen, the installation wizard will point to the latest parcels of CDH available. To specify CDH 5.1.3 as an option for parcel installation, click **More Options** and add the following repository URL to the list of parcel repositories:

`http://archive.cloudera.com/cdh5/parcels/5.1.3/`. The screen will refresh after a few seconds, and you can choose CDH 5.1.3.

On the installation wizard's **Cluster Setup** page, choose **Custom Services**, and choose the services you want installed in the cluster. Be sure to choose **Isilon** among the selected services, do not select the **HDFS** or **Impala** services, and do not check **Include Cloudera Navigator** at the bottom of the **Cluster Setup** page. Also, on the **Role Assignments** page, be sure to specify the hosts that will serve as gateway roles for the Isilon service. You can add gateway roles to one, some, or all nodes in the cluster.

### Installing a Secure Cluster with Isilon

To set up a secure cluster with Isilon using Kerberos, perform the following steps:

1. Create an unsecure Cloudera Manager cluster as described above in [Installing Cloudera Manager with the Isilon Storage Service](#).
2. Follow the Isilon documentation to enable Kerberos for your access zone: <https://community.emc.com/docs/DOC-39522>. This includes adding a Kerberos authentication provider to your Isilon access zone.
3. Add the following proxy users in Isilon if your Cloudera Manager cluster includes the corresponding CDH services. The procedure for configuring proxy users is described in the Isilon documentation, <https://community.emc.com/docs/DOC-39522>.
  - proxy user `hdfs` for `hdfs` user.
  - proxy user `mapred` for `mapred` user.
  - proxy user `hive` for `hive` user.
  - proxy user `oozie` for `oozie` user
  - proxy user `flume` for `flume` user
  - proxy user `hue` for `hue` user
4. Follow the Cloudera Manager documentation for information on configuring a secure cluster with Kerberos: [Configuring Authentication in Cloudera Manager](#).


## The Key-Value Store Indexer Service

The Key-Value Store Indexer service uses the [Lily HBase Indexer Service](#) to index the stream of records being added to HBase tables. Indexing allows you to query data stored in HBase with the [Solr service](#).

The Key-Value Store Indexer service is installed in the same parcel or package along with the CDH 5 or Solr service.

### Adding the Key-Value Store Indexer Service

Required Role: **Cluster Administrator** **Full Administrator**

1. On the Home page, click  to the right of the cluster name and select **Add a Service**. A list of service types display. You can add one type of service at a time.
2. Select the **Key-Value Store Indexer** service and click **Continue**.
3. Select the radio button next to the services on which the new service should depend and click **Continue**.
4. Customize the assignment of role instances to hosts. The wizard evaluates the hardware configurations of the hosts to determine the best hosts for each role. The wizard assigns all worker roles to the same set of hosts to which the HDFS DataNode role is assigned. These assignments are typically acceptable, but you can reassign them if necessary.

Click a field below a role to display a dialog containing a list of hosts. If you click a field containing multiple hosts, you can also select **All Hosts** to assign the role to all hosts or **Custom** to display the pageable hosts dialog.

The following shortcuts for specifying hostname patterns are supported:

- Range of hostnames (without the domain portion)

Range Definition	Matching Hosts
10.1.1.[1-4]	10.1.1.1, 10.1.1.2, 10.1.1.3, 10.1.1.4
host[1-3].company.com	host1.company.com, host2.company.com, host3.company.com
host[07-10].company.com	host07.company.com, host08.company.com, host09.company.com, host10.company.com

- IP addresses
- Rack name

## Configuring CDH and Managed Services

Click the **View By Host** button for an overview of the role assignment by hostname ranges.

5. Click **Continue**.
6. Review the configuration changes to be applied. Confirm the settings entered for file system paths. The file paths required vary based on the services to be installed.

■ **Warning:** DataNode data directories should not be placed on NAS devices.

Click **Continue**. The wizard starts the services.

7. Click **Continue**.
8. Click **Finish**.

### Enabling Morphlines with Search and HBase Indexing

**Required Role:** **Configurator** **Cluster Administrator** **Full Administrator**

Cloudera Morphlines is an open source framework that reduces the time and skills necessary to build or change Search indexing applications. A morphline is a rich configuration file that simplifies defining an ETL transformation chain.

1. Go to the Indexer service.
2. Click the **Configuration** tab.
3. Create the necessary configuration files, and modify the content in the following properties under the **Service-Wide > Morphlines** category:
  - **Morphlines File** — Text that goes into the `morphlines.conf` used by HBase indexers. You should use `$ZK_HOST` in this file instead of specifying a ZooKeeper quorum. Cloudera Manager automatically replaces the `$ZK_HOST` variable with the correct value during the Solr configuration deployment.
  - **Custom MIME-types File** — Text that goes verbatim into the `custom-mimetypes.xml` file used by HBase Indexers with the `detectMimeTypes` command. See the [Cloudera Morphlines Reference Guide](#) for details on this command.
  - **Grok Dictionary File** — Text that goes verbatim into the `grok-dictionary.conf` file used by HBase Indexers with the `grok` command. See the [Cloudera Morphlines Reference Guide](#) for details of this command.

See [Extracting, Transforming, and Loading Data With Cloudera Morphlines](#) for information about using morphlines with Search and HBase.

### The MapReduce and YARN Services

CDH supports two versions of the MapReduce computation framework: MRv1 and MRv2, which are implemented by the [MapReduce](#) (MRv1) and [YARN](#) (MRv2) services. YARN is backwards-compatible with MapReduce. (All jobs that run against MapReduce will also run in a YARN cluster).

The MRv2 YARN architecture splits the two primary responsibilities of the JobTracker — resource management and job scheduling/monitoring — into separate daemons: a global ResourceManager (RM) and per-application ApplicationMasters (AM). With MRv2, the ResourceManager (RM) and per-node NodeManagers (NM) form the data-computation framework. The ResourceManager service effectively replaces the functions of the JobTracker, and NodeManagers run on worker hosts instead of TaskTracker daemons. The per-application ApplicationMaster is, in effect, a framework-specific library and negotiates resources from the ResourceManager and works with the NodeManagers to execute and monitor the tasks. For details of this architecture, see [Apache Hadoop NextGen MapReduce \(YARN\)](#).

- The Cloudera Manager Admin Console has different methods for displaying MapReduce and YARN job history. See [MapReduce Jobs](#) and [YARN Applications](#).
- For information on configuring the MapReduce and YARN services for high availability, see [MapReduce \(MRv1\) and YARN \(MRv2\) High Availability](#) on page 237

- For information on configuring MapReduce and YARN resource management features, see [Resource Management](#) on page 171.

### Defaults and Recommendations

- In a Cloudera Manager deployment of a CDH 4 cluster, the MapReduce service is the default MapReduce computation framework. You can create a YARN service in a CDH 4 cluster, but it is not considered production ready.
- In a Cloudera Manager deployment of a CDH 5 cluster, the YARN service is the default MapReduce computation framework. In CDH 5, the MapReduce service has been deprecated. However, the MapReduce service is fully supported for backward compatibility through the CDH 5 life cycle.
- For production uses, Cloudera recommends that *only one* MapReduce framework should be running at any given time. If development needs or other use case requires switching between MapReduce and YARN, both services can be configured at the same time, but only one should be in a running (to fully optimize the hardware resources available).

### Migrating from MapReduce to YARN

Cloudera Manager provides a wizard described in [Importing MapReduce Configurations to YARN](#) on page 83 to easily migrate MapReduce configurations to YARN. The wizard performs all the steps ([Switching Between MapReduce and YARN Services](#) on page 79, [Updating Dependent Services](#) on page 79, and [Configuring Alternatives Priority](#) on page 80) on this page.

For detailed information on migrating from MapReduce to YARN, see [Migrating from MapReduce 1 \(MRv1\) to MapReduce 2 \(MRv2, YARN\)](#).

### Switching Between MapReduce and YARN Services

**Required Role:** Configurator Cluster Administrator Full Administrator

MapReduce and YARN use separate sets of configuration files. No files are removed or altered when you change to a different framework. To change from YARN to MapReduce (or vice versa):

1. (Optional) Configure the new MapReduce or YARN framework service.
2. [Update dependent services](#) to use the chosen framework.
3. Configure the [alternatives priority](#).
4. [Redeploy the Oozie ShareLib](#).
5. Redeploy the client configuration.
6. Start the framework service to switch to.
7. (Optional) Stop the unused framework service to free up the resources it uses.

### Updating Dependent Services

**Required Role:** Configurator Cluster Administrator Full Administrator

When you change the MapReduce framework, the dependent services that must be updated to use the new framework are:

- Hive
- Sqoop 2
- Oozie

To update a service:

1. Go to the service.
2. Click the **Configuration** tab.
3. Select **Service-Wide**.
4. Click the **MapReduce Service** property and select the YARN or MapReduce service.
5. Click **Save Changes** to commit the changes.

### 6. Select **Actions** > **Restart**.

The Hue service is automatically reconfigured to use the same framework as Oozie and Hive. This cannot be changed. To update the Hue service:

1. Go to the Hue service.
2. Select **Actions** > **Restart**.

### Configuring Alternatives Priority

**Required Role:** **Configurator** **Cluster Administrator** **Full Administrator**

The alternatives priority property determines which service—MapReduce or YARN—is used by clients to run MapReduce jobs. The service with a higher value of the property is used. In CDH 4, the MapReduce service alternatives priority is set to 92 and the YARN service is set to 91. In CDH 5, the values are reversed; the MapReduce service alternatives priority is set to 91 and the YARN service is set to 92.

To configure the alternatives priority:

1. Go to the MapReduce or YARN service.
2. Click the **Configuration** tab.
3. Expand the **Gateway Default Group** node.
4. In the **Alternatives Priority** property, set the priority value.
5. Click **Save Changes** to commit the changes.
6. Redeploy the client configuration.

### The MapReduce Service

For an overview of computation frameworks, and their usage and restrictions, and common tasks, see [The MapReduce and YARN Services](#) on page 78.

### Configuring the MapReduce Scheduler

**Required Role:** **Configurator** **Cluster Administrator** **Full Administrator**

The MapReduce service is configured by default to use the FairScheduler. You can change the scheduler type to FIFO or Capacity Scheduler. You can also modify the Fair Scheduler and Capacity Scheduler configuration. For further information on schedulers, see [Schedulers](#) on page 171.

### Configuring the Task Scheduler Type

1. Go to the MapReduce service.
2. Click the **Configuration** tab.
3. Expand the **JobTracker Default Group** category and click the **Classes** category.
4. Click the **Value** field of the Task Scheduler row and select a scheduler.
5. Click **Save Changes** to commit the changes.
6. Restart the JobTracker to apply the new configuration:
  - a. Click the **Instances** tab.
  - b. Click the **JobTracker** role.
  - c. Select **Actions for Selected** > **Restart**.

### Modifying the Scheduler Configuration

1. Go to the MapReduce service.
2. Click the **Configuration** tab.
3. Click the **Jobs** subcategory of the **JobTracker Default Group** category.
4. Click a property and modify the configuration.
5. Click **Save Changes** to commit the changes.



6. Restart the JobTracker to apply the new configuration:
  - a. Click the **Instances** tab.
  - b. Click the **JobTracker** role.
  - c. Select **Actions for Selected** > **Restart**.

### Configuring the MapReduce Service to Save Job History

**Required Role:** Configurator Cluster Administrator Full Administrator

Normally job history is saved on the host on which the JobTracker is running. You can configure JobTracker to write information about every job that completes to a specified HDFS location. By default, the information is retained for 7 days.

### Enabling Map Reduce Job History To Be Saved to HDFS

1. Create a folder in HDFS to contain the history information. When creating the folder, set the owner and group to `mapred:hadoop` with permission setting 775.
2. Go to the MapReduce service.
3. Click the **Configuration** tab.
4. Expand the **JobTracker Default Group** category and click the **Paths** subcategory.
5. Set the **Completed Job History Location** property to the location that you created in [step 1](#).
6. Click **Save Changes**.
7. Restart the MapReduce service.

### Setting the Job History Retention Duration

1. Select the **JobTracker Default Group** category.
2. Set the **Job History Files Maximum Age** property (`mapreduce.jobhistory.max-age-ms`) to the length of time (in milliseconds, seconds, minutes, or hours) that you want job history files to be kept.
3. Restart the MapReduce service.

The Job History Files Cleaner runs at regular intervals to check for job history files that are ready to be deleted. By default, the interval is 24 hours. To change the frequency with which the Job History Files Cleaner runs:

1. Select the **JobTracker Default Group** category.
2. Set the **Job History Files Cleaner Interval** property (`mapreduce.jobhistory.cleaner.interval`) to the desired frequency (in milliseconds, seconds, minutes, or hours).
3. Restart the MapReduce service.

### Configuring Client Overrides

A configuration property qualified with **(Client Override)** is a server-side setting that ignores any value a client tries to set for that property. It performs the same role as its unqualified counterpart, and applies the configuration to the service with the setting `<final>true</final>`.


For example, if you set the Map task heap property to 1 GB in the job configuration code, but the service's heap property qualified with (Client Override) is set to 500 MB, then 500 MB is applied.

### The YARN Service

For an overview of computation frameworks, and their usage and restrictions, and common tasks, see [The MapReduce and YARN Services](#) on page 78.

### Adding the YARN Service

**Required Role:** Cluster Administrator Full Administrator

1. On the Home page, click  to the right of the cluster name and select **Add a Service**. A list of service types display. You can add one type of service at a time.

- 2. Click the **YARN (MR2 Included)** radio button and click **Continue**.
- 3. Select the radio button next to the services on which the new service should depend and click **Continue**.
- 4. Customize the assignment of role instances to hosts. The wizard evaluates the hardware configurations of the hosts to determine the best hosts for each role. The wizard assigns all worker roles to the same set of hosts to which the HDFS DataNode role is assigned. These assignments are typically acceptable, but you can reassign them if necessary.

Click a field below a role to display a dialog containing a list of hosts. If you click a field containing multiple hosts, you can also select **All Hosts** to assign the role to all hosts or **Custom** to display the pageable hosts dialog.

The following shortcuts for specifying hostname patterns are supported:

- Range of hostnames (without the domain portion)

Range Definition	Matching Hosts
10.1.1.[1-4]	10.1.1.1, 10.1.1.2, 10.1.1.3, 10.1.1.4
host[1-3].company.com	host1.company.com, host2.company.com, host3.company.com
host[07-10].company.com	host07.company.com, host08.company.com, host09.company.com, host10.company.com

- IP addresses
- Rack name

Click the **View By Host** button for an overview of the role assignment by hostname ranges.

Configuring Memory Settings for YARN and MRv2

The memory configuration for YARN and MRv2 memory is important to get the best performance from your cluster. Several different settings are involved. The table below shows the default settings, as well as the settings that Cloudera recommends, for each configuration option. See [The MapReduce and YARN Services](#) on page 78 for more configuration specifics and, for detailed tuning advice with sample configurations, see [Tuning the Cluster for MapReduce v2 \(YARN\)](#).

Table 1: YARN and MRv2 Memory Configuration

Cloudera Manager Property Name	CDH Property Name	Default Configuration	Cloudera Tuning Guidelines
Container Memory Minimum	yarn.scheduler.minimum-allocation-mb	1 GB	0
Container Memory Maximum	yarn.scheduler.maximum-allocation-mb	64 GB	amount of memory on largest node
Container Memory Increment	yarn.scheduler.increment-allocation-mb	512 MB	Use a fairly large value, such as 128 MB
Container Memory	yarn.nodemanager.resource.memory-mb	8 GB	8 GB
Map Task Memory	mapreduce.map.memory.mb	1 GB	1 GB
Reduce Task Memory	mapreduce.reduce.memory.mb	1 GB	1 GB
Map Task Java Opts Base	mapreduce.map.java.opts	-Djava.net.preferIPv4Stack=true	-Djava.net.preferIPv4Stack=true -Xmx768m
Reduce Task Java Opts Base	mapreduce.reduce.java.opts	-Djava.net.preferIPv4Stack=true	-Djava.net.preferIPv4Stack=true -Xmx768m
ApplicationMaster Memory	yarn.app.mapreduce.am.resource.mb	1 GB	1 GB

Cloudera Manager Property Name	CDH Property Name	Default Configuration	Cloudera Tuning Guidelines
ApplicationMaster Java Opts Base	yarn.app.mapreduce.am.command-opts	-Djava.net.preferIPv4Stack=true	-Djava.net.preferIPv4Stack=true -Xmx768m

### Configuring Directories

**Required Role:** Cluster Administrator Full Administrator

#### Creating the Job History Directory

When adding the YARN service, the **Add Service** wizard automatically creates a job history directory. If you quit the **Add Service** wizard or it does not finish, you can create the directory outside the wizard:

1. Go to the YARN service.
2. Select **Actions** > **Create Job History Dir.**
3. Click **Create Job History Dir** again to confirm.

#### Creating the NodeManager Remote Application Log Directory

When adding the YARN service, the **Add Service** wizard automatically creates a remote application log directory. If you quit the **Add Service** wizard or it does not finish, you can create the directory outside the wizard:


1. Go to the YARN service.
2. Select **Actions** > **Create NodeManager Remote Application Log Directory.**
3. Click **Create NodeManager Remote Application Log Directory** again to confirm.

### Importing MapReduce Configurations to YARN

**Required Role:** Cluster Administrator Full Administrator

- **Warning:** In addition to importing configuration settings, the import process:
  - Configures services to use YARN as the MapReduce computation framework instead of MapReduce.
  - Overwrites existing YARN configuration and role assignments.

When you upgrade from CDH 4 to CDH 5, you can import MapReduce configurations to YARN as part of the upgrade wizard. If you do not import configurations during upgrade, you can manually import the configurations at a later time:

1. Go to the YARN service page.
2. Stop the YARN service.
3. Select **Actions** > **Import MapReduce Configuration**. The import wizard presents a warning letting you know that it will import your configuration, restart the YARN service and its dependent services, and update the client configuration.
4. Click **Continue** to proceed. The next page indicates some additional configuration required by YARN.
5. Verify or modify the configurations and click **Continue**. The Switch Cluster to MR2 step proceeds.
6. When all steps have been completed, click **Finish**.
7. (Optional) Remove the MapReduce service.
  - a. Click the **Home** tab.
  - b. In the MapReduce row, right-click  and select **Delete**. Click **Delete** to confirm.
8. Recompile JARs used in MapReduce applications. For further information, see [For MapReduce Programmers: Writing and Running Jobs](#).
- 9.

## Configuring CDH and Managed Services

### Configuring the YARN Scheduler

Required Role: **Configurator** **Cluster Administrator** **Full Administrator**

The YARN service is configured by default to use the FairScheduler. You can change the scheduler type to FIFO or Capacity Scheduler. You can also modify the Fair Scheduler and Capacity Scheduler configuration. For further information on schedulers, see [Schedulers](#) on page 171.

### Configuring the Scheduler Type

1. Go to the YARN service.
2. Click the **Configuration** tab.
3. Expand the **ResourceManager Default Group** and click the **Scheduler Class** property.
4. Select a scheduler class.
5. Click **Save Changes** to commit the changes.
6. Restart the YARN service.

### Modifying the Scheduler Configuration

1. Go to the YARN service.
2. Click the **Configuration** tab.
3. Click the **ResourceManager Default Group** category.
4. Click a property and modify the configuration.
5. Click **Save Changes** to commit the changes.
6. Restart the YARN service.

### Dynamic Resource Management

In addition to the [static resource management](#) available to all services, the YARN service also supports dynamic management of its static allocation. See [Dynamic Resource Pools](#) on page 177.

### Configuring YARN for Long-running Applications

On a secure cluster, long-running applications such as Spark Streaming jobs will need additional configuration since the default settings only allow the `hdfs` user's delegation tokens a maximum lifetime of 7 days, which is not always sufficient. For instructions on how to work around this issue, see [Configuring YARN for Long-running Applications](#).

## The Oozie Service

Cloudera Manager installs the Oozie service as part of the CDH installation.

You can elect to have the service created and started as part of the Installation wizard. If you elect not to create the service using the installation wizard, you can use the **Add Service** wizard to perform the installation. The wizard will automatically configure and start the dependent services and the Oozie service. See [Adding a Service](#) on page 30 for instructions.

For information on configuring Oozie for high availability, see [Oozie High Availability](#) on page 262.

### Redeploying the Oozie ShareLib

Required Role: **Configurator** **Cluster Administrator** **Full Administrator**

Some Oozie actions – specifically DistCp, Streaming, Pig, Sqoop, and Hive – require external JAR files in order to run. Instead of having to keep these JAR files in each workflow's `lib` folder, or forcing you to manually manage them via the `oozie.libpath` property on every workflow using one of these actions, Oozie provides the ShareLib. The ShareLib behaves very similarly to `oozie.libpath`, except that it's specific to the aforementioned actions and their required JARs.

When you upgrade CDH or [switch between MapReduce and YARN](#) computation frameworks, redeploy the Oozie ShareLib as follows:

1. Go to the Oozie service.
2. Select **Actions** > **Stop**.
3. Select **Actions** > **Install Oozie ShareLib**.
4. Select **Actions** > **Start**.

#### Adding Schema to Oozie

**Required Role:** **Configurator** **Cluster Administrator** **Full Administrator**

For CDH 4.x Cloudera Manager configures Oozie to recognize only the schema available in CDH 4.0.0, even though more were added later. If you want to use any additional schema, do the following:

1. In the Cloudera Manager Admin Console, go to the Oozie service.
2. Click the **Configuration** tab.
3. Click **Oozie Server Default Group**.
4. Select the **Oozie SchemaService Workflow Extension Schemas** property.
5. Enter the desired schema from [Table 2: Oozie Schema - CDH 5](#) on page 85 or [Table 3: Oozie Schema - CDH 4](#) on page 86, appending `.xsd` to each entry.
6. Click **Save Changes** to commit the changes.
7. Restart the Oozie service.

■ **Note:** Releases are only included in the following tables if there was a schema added or removed. If a release is not in the table, it should have the same set of schemas as the previous release that is in the table.

**Table 2: Oozie Schema - CDH 5**

	CDH 5.2.0	CDH 5.1.0	CDH 5.0.1	CDH 5.0.0
<b>distcp</b>	distcp-action-0.1 distcp-action-0.2	distcp-action-0.1 distcp-action-0.2	distcp-action-0.1 distcp-action-0.2	distcp-action-0.1 distcp-action-0.2
<b>email</b>	email-action-0.1	email-action-0.1 email-action-0.2	email-action-0.1	email-action-0.1
<b>hive</b>	hive-action-0.2 hive-action-0.3 hive-action-0.4 hive-action-0.5	hive-action-0.2 hive-action-0.3 hive-action-0.4 hive-action-0.5	hive-action-0.2 hive-action-0.3 hive-action-0.4 hive-action-0.5	hive-action-0.2 hive-action-0.3 hive-action-0.4 hive-action-0.5
<b>HiveServer2</b>	hive2-action-0.1			
<b>oozie-bundle</b>	oozie-bundle-0.1 oozie-bundle-0.2	oozie-bundle-0.1 oozie-bundle-0.2	oozie-bundle-0.1 oozie-bundle-0.2	oozie-bundle-0.1 oozie-bundle-0.2
<b>oozie-coordinator</b>	oozie-coordinator-0.1 oozie-coordinator-0.2 oozie-coordinator-0.3 oozie-coordinator-0.4	oozie-coordinator-0.1 oozie-coordinator-0.2 oozie-coordinator-0.3 oozie-coordinator-0.4	oozie-coordinator-0.1 oozie-coordinator-0.2 oozie-coordinator-0.3 oozie-coordinator-0.4	oozie-coordinator-0.1 oozie-coordinator-0.2 oozie-coordinator-0.3 oozie-coordinator-0.4
<b>oozie-sla</b>	oozie-sla-0.1 oozie-sla-0.2	oozie-sla-0.1 oozie-sla-0.2	oozie-sla-0.1 oozie-sla-0.2	oozie-sla-0.1 oozie-sla-0.2

	CDH 5.2.0	CDH 5.1.0	CDH 5.0.1	CDH 5.0.0
<b>oozie-workflow</b>	oozie-workflow-0.1 oozie-workflow-0.2 oozie-workflow-0.2.5 oozie-workflow-0.3 oozie-workflow-0.4 oozie-workflow-0.4.5 oozie-workflow-0.5	oozie-workflow-0.1 oozie-workflow-0.2 oozie-workflow-0.2.5 oozie-workflow-0.3 oozie-workflow-0.4 oozie-workflow-0.4.5 oozie-workflow-0.5	oozie-workflow-0.1 oozie-workflow-0.2 oozie-workflow-0.2.5 oozie-workflow-0.3 oozie-workflow-0.4 oozie-workflow-0.4.5 oozie-workflow-0.5	oozie-workflow-0.1 oozie-workflow-0.2 oozie-workflow-0.2.5 oozie-workflow-0.3 oozie-workflow-0.4 oozie-workflow-0.5
<b>shell</b>	shell-action-0.1 shell-action-0.2 shell-action-0.3	shell-action-0.1 shell-action-0.2 shell-action-0.3	shell-action-0.1 shell-action-0.2 shell-action-0.3	shell-action-0.1 shell-action-0.2 shell-action-0.3
<b>sqoop</b>	sqoop-action-0.2 sqoop-action-0.3 sqoop-action-0.4	sqoop-action-0.2 sqoop-action-0.3 sqoop-action-0.4	sqoop-action-0.2 sqoop-action-0.3 sqoop-action-0.4	sqoop-action-0.2 sqoop-action-0.3 sqoop-action-0.4
<b>ssh</b>	ssh-action-0.1 ssh-action-0.2	ssh-action-0.1 ssh-action-0.2	ssh-action-0.1 ssh-action-0.2	ssh-action-0.1 ssh-action-0.2

Table 3: Oozie Schema - CDH 4

	CDH 4.3.0	CDH 4.2.0	CDH 4.1.0	CDH 4.0.0
<b>distcp</b>	distcp-action-0.1 distcp-action-0.2	distcp-action-0.1 distcp-action-0.2	distcp-action-0.1	distcp-action-0.1
<b>email</b>	email-action-0.1	email-action-0.1	email-action-0.1	email-action-0.1
<b>hive</b>	hive-action-0.2 hive-action-0.3 hive-action-0.4 hive-action-0.5	hive-action-0.2 hive-action-0.3 hive-action-0.4	hive-action-0.2 hive-action-0.3 hive-action-0.4	hive-action-0.2
<b>oozie-bundle</b>	oozie-bundle-0.1 oozie-bundle-0.2	oozie-bundle-0.1 oozie-bundle-0.2	oozie-bundle-0.1 oozie-bundle-0.2	oozie-bundle-0.1
<b>oozie-coordinator</b>	oozie-coordinator-0.1 oozie-coordinator-0.2 oozie-coordinator-0.3 oozie-coordinator-0.4	oozie-coordinator-0.1 oozie-coordinator-0.2 oozie-coordinator-0.3 oozie-coordinator-0.4	oozie-coordinator-0.1 oozie-coordinator-0.2 oozie-coordinator-0.3 oozie-coordinator-0.4	oozie-coordinator-0.1 oozie-coordinator-0.2 oozie-coordinator-0.3
<b>oozie-sla</b>	oozie-sla-0.1	oozie-sla-0.1	oozie-sla-0.1	oozie-sla-0.1
<b>oozie-workflow</b>	oozie-workflow-0.1 oozie-workflow-0.2 oozie-workflow-0.2.5	oozie-workflow-0.1 oozie-workflow-0.2 oozie-workflow-0.2.5	oozie-workflow-0.1 oozie-workflow-0.2 oozie-workflow-0.2.5	oozie-workflow-0.1 oozie-workflow-0.2 oozie-workflow-0.2.5

	CDH 4.3.0	CDH 4.2.0	CDH 4.1.0	CDH 4.0.0
	oozie-workflow-0.3 oozie-workflow-0.4 oozie-workflow-0.4.5	oozie-workflow-0.3 oozie-workflow-0.4	oozie-workflow-0.3 oozie-workflow-0.4	oozie-workflow-0.3
<b>shell</b>	shell-action-0.1 shell-action-0.2 shell-action-0.3	shell-action-0.1 shell-action-0.2 shell-action-0.3	shell-action-0.1 shell-action-0.2 shell-action-0.3	shell-action-0.1
<b>sqoop</b>	sqoop-action-0.2 sqoop-action-0.3 sqoop-action-0.4	sqoop-action-0.2 sqoop-action-0.3 sqoop-action-0.4	sqoop-action-0.2 sqoop-action-0.3 sqoop-action-0.4	sqoop-action-0.2
<b>ssh</b>	ssh-action-0.1 ssh-action-0.2	ssh-action-0.1	ssh-action-0.1	ssh-action-0.1

### Enabling the Oozie Web Console

Required Role: **Configurator** **Cluster Administrator** **Full Administrator**

1. Download [ext-2.2](#). Extract the contents of the file to `/var/lib/oozie/` on the same host as the Oozie Server.
2. In the Cloudera Manager Admin Console, go to the Oozie service.
3. Click the **Configuration** tab.
4. Check **Enable Oozie server web console**.
5. Click **Save Changes** to commit the changes.
6. Restart the Oozie service.

### Using an External Database for Oozie

The default database for Oozie is Derby. Cloudera recommends that you use a production database instead, for the following reasons:

- Derby runs in embedded mode and it is not possible to monitor its health.
- It is not clear how to implement a live backup strategy for the embedded Derby database, though it may be possible.
- Under load, Cloudera has observed locks and rollbacks with the embedded Derby database which don't happen with server-based databases.

The databases that Oozie supports are listed at:

- [CDH 4 supported databases](#)
- [CDH 5 supported databases](#)

### Setting up an External Database for Oozie

See [PostgreSQL](#) on page 87, [MySQL](#) on page 88, or [Oracle](#) on page 89 for the procedure for setting up an Oozie database.

#### PostgreSQL

Use the procedure that follows to configure Oozie to use PostgreSQL instead of Apache Derby.

Install PostgreSQL 8.4.x or 9.0.x.

See [External PostgreSQL Database](#).

## Configuring CDH and Managed Services

Create the Oozie user and Oozie database.

For example, using the PostgreSQL `psql` command-line tool:

```
$ psql -U postgres
Password for user postgres: *****

postgres=# CREATE ROLE oozie LOGIN ENCRYPTED PASSWORD 'oozie'
NOSUPERUSER INHERIT CREATEDB NOCREATEROLE;
CREATE ROLE

postgres=# CREATE DATABASE "oozie" WITH OWNER = oozie
ENCODING = 'UTF8'
TABLESPACE = pg_default
LC_COLLATE = 'en_US.UTF8'
LC_CTYPE = 'en_US.UTF8'
CONNECTION LIMIT = -1;
CREATE DATABASE

postgres=# \q
```

Configure PostgreSQL to accept network connections for the oozie user.

1. Edit the `postgresql.conf` file and set the `listen_addresses` property to `*`, to make sure that the PostgreSQL server starts listening on all your network interfaces. Also make sure that the `standard_conforming_strings` property is set to `off`.
2. Edit the PostgreSQL `data/pg_hba.conf` file as follows:

```
host      oozie          oozie          0.0.0.0/0          md5
```

Reload the PostgreSQL configuration.

```
$ sudo -u postgres pg_ctl reload -s -D /opt/PostgreSQL/8.4/data
```

### MySQL

Use the procedure that follows to configure Oozie to use MySQL instead of Apache Derby.

Install and start MySQL 5.x

See [MySQL Database](#).

Create the Oozie database and Oozie MySQL user.

For example, using the MySQL `mysql` command-line tool:

```
$ mysql -u root -p
Enter password: *****

mysql> create database oozie;
Query OK, 1 row affected (0.03 sec)

mysql> grant all privileges on oozie.* to 'oozie'@'localhost' identified by 'oozie';
Query OK, 0 rows affected (0.03 sec)

mysql> grant all privileges on oozie.* to 'oozie'@'%' identified by 'oozie';
Query OK, 0 rows affected (0.03 sec)

mysql> exit
Bye
```



[Add the MySQL JDBC driver JAR to Oozie.](#)

Copy or symbolically link the MySQL JDBC driver JAR into the `/var/lib/oozie/` directory.

- **Note:** You must manually download the MySQL JDBC driver JAR file.

## Oracle

Use the procedure that follows to configure Oozie to use Oracle 11g instead of Apache Derby.

[Install and start Oracle 11g.](#)

Use [Oracle's instructions](#).

[Create the Oozie Oracle user and grant privileges.](#)

The following example uses the Oracle `sqlplus` command-line tool, and shows the privileges Cloudera recommends.

```
$ sqlplus system@localhost

Enter password: *****

SQL> create user oozie identified by oozie default tablespace users temporary tablespace
temp;

User created.

SQL> grant alter any index to oozie;
grant alter any table to oozie;
grant alter database link to oozie;
grant create any index to oozie;
grant create any sequence to oozie;
grant create database link to oozie;
grant create session to oozie;
grant create table to oozie;
grant drop any sequence to oozie;
grant select any dictionary to oozie;
grant drop any table to oozie;
grant create procedure to oozie;
grant create trigger to oozie;

SQL> exit

$
```

- **Important:**

Do *not* make the following grant:

```
grant select any table;
```

[Add the Oracle JDBC driver JAR to Oozie.](#)

Copy or symbolically link the Oracle JDBC driver JAR into the `/var/lib/oozie/` directory.

- **Note:** You must manually download the Oracle JDBC driver JAR file.

## Creating the Oozie Database Schema

After configuring Oozie database information and creating the corresponding database, create the Oozie database schema. Oozie provides a database tool for this purpose.

- **Note:** The Oozie database tool uses Oozie configuration files to connect to the database to perform the schema creation; before you use the tool, make you have created a database and configured Oozie to work with it as described above.

The Oozie database tool works in 2 modes: it can create the database, or it can produce an SQL script that a database administrator can run to create the database manually. If you use the tool to create the database schema, you must have the permissions needed to execute DDL operations.

### To run the Oozie database tool against the database

- **Important:** This step must be done as the `oozie` Unix user, otherwise Oozie may fail to start or work properly because of incorrect file permissions.

```
$ sudo -u oozie /usr/lib/oozie/bin/ooziedb.sh create -run
```

You should see output such as the following (the output of the script may differ slightly depending on the database vendor) :

```
Validate DB Connection.
DONE
Check DB schema does not exist
DONE
Check OOZIE_SYS table does not exist
DONE
Create SQL schema
DONE
Create OOZIE_SYS table
DONE

Oozie DB has been created for Oozie version '4.0.0-cdh5.0.0'

The SQL commands have been written to: /tmp/ooziedb-5737263881793872034.sql
```

### To create the upgrade script

- **Important:** This step must be done as the `oozie` Unix user, otherwise Oozie may fail to start or work properly because of incorrect file permissions.

Run `/usr/lib/oozie/bin/ooziedb.sh create -sqlfile SCRIPT`. For example:

```
$ sudo -u oozie /usr/lib/oozie/bin/ooziedb.sh create -sqlfile oozie-create.sql
```

You should see output such as the following (the output of the script may differ slightly depending on the database vendor) :

```
Validate DB Connection.
DONE
Check DB schema does not exist
DONE
Check OOZIE_SYS table does not exist
DONE
Create SQL schema
DONE
Create OOZIE_SYS table
DONE

Oozie DB has been created for Oozie version '4.0.0-cdh5.0.0'

The SQL commands have been written to: oozie-create.sql
```

WARN: The SQL commands have NOT been executed, you must use the '-run' option

- **Important:** If you used the `-sqlfile` option instead of `-run`, Oozie database schema has not been created. You must run the `oozie-create.sql` script against your database.

### Configuring Oozie to Use an External Database

Required Role: **Configurator** **Cluster Administrator** **Full Administrator**

1. In the Cloudera Manager Admin Console, go to the Oozie service.
2. Click the **Configuration** tab.
3. Expand **Oozie Server Default Group** and click **Database**.
4. Specify the settings for **Oozie Server database type**, **Oozie Server database name**, **Oozie Server database host**, **Oozie Server database user**, and **Oozie Server database password**.
5. Click **Save Changes** to commit the changes.
6. Restart the Oozie service.

### The Solr Service

You can install the Solr service through the Cloudera Manager installation wizard, using either parcels or packages. See [Installing Search](#).

You can elect to have the service created and started as part of the Installation wizard. If you elect not to create the service using the Installation wizard, you can use the **Add Service** wizard to perform the installation. The wizard will automatically configure and start the dependent services and the Solr service. See [Adding a Service](#) on page 30 for instructions.

For further information on the Solr service, see:

- **CDH 4** - [Cloudera Search Documentation for CDH 4](#)
- **CDH 5** - [Cloudera Search Guide](#)

The following sections describe how to configure other CDH components to work with the Solr service.

### Configuring the Flume Morphline Solr Sink for Use with the Solr Service

Required Role: **Configurator** **Cluster Administrator** **Full Administrator**

To use a Flume Morphline Solr sink, the Flume service must be running on your cluster. See the [Flume Near Real-Time Indexing Reference \(CDH 4\)](#) or [Flume Near Real-Time Indexing Reference \(CDH 5\)](#) or information about the Flume Morphline Solr Sink and [The Flume Service](#) on page 44.

1. Go to the Flume service.
2. Click the **Configuration** tab.
3. Expand the **Agent** default role group and click the **Flume-NG Solr Sink** category.
4. Edit the following settings, which are templates that you must modify for your deployment:
  - **Morphlines File** (`morphlines.conf`) - Configures Morphlines for Flume agents. You must use `$ZK_HOST` in this field instead of specifying a ZooKeeper quorum. Cloudera Manager automatically replaces the `$ZK_HOST` variable with the correct value during the Flume configuration deployment.
  - **Custom MIME-types File** (`custom-mimetypes.xml`) - Configuration for the `detectMimeTypes` command. See the [Cloudera Morphlines Reference Guide](#) for details on this command.
  - **Grok Dictionary File** (`grok-dictionary.conf`) - Configuration for the `grok` command. See the [Cloudera Morphlines Reference Guide](#) for details on this command.

Once configuration is complete, Cloudera Manager automatically deploys the required files to the Flume agent process directory when it starts the Flume agent. Therefore, you can reference the files in the [Flume agent](#)

## Configuring CDH and Managed Services

[configuration](#) using their relative path names. For example, you can use the name `morphlines.conf` to refer to the location of the Morphlines configuration file.

## Deploying Solr with Hue

**Required Role:** Configurator Cluster Administrator Full Administrator

In CDH 4.3 and earlier, in order to use Solr with Hue, you must update the URL for the Solr Server in the Hue Server advanced configuration snippet.

1. Go to the **Hue** service.
2. Click the **Configuration** tab.
3. Search for the word "snippet". This will display a set of Hue advanced configuration snippet properties.
4. Add information about your Solr host to the **Hue Server Configuration Advanced Configuration Snippet for hue\_safety\_valve\_server.ini** found under the **Hue Server Default Group > Advanced** category. For example, if your hostname is `SOLR_HOST`, you might add the following:

```
[search]
## URL of the Solr Server
solr_url=http://SOLR_HOST:8983/solr
```

5. Click **Save Changes** to save your advanced configuration snippet changes.
6. Restart the Hue Service.

- **Important:** If you are using parcels with CDH 4.3, you must register the "hue-search" application manually or access will fail. You do not need to do this if you are using CDH 4.4 or later.

1. Stop the Hue service.

2. From the command line do the following:

- ```
a. cd /opt/cloudera/parcels/CDH 4.3.0-1.cdh4.3.0.pXXX/share/hue
```

(Substitute your own local repository path for the `/opt/cloudera/parcels/...` if yours is different, and specify the appropriate name of the CDH 4.3 parcel that exists in your repository.)

- ```
b. ./build/env/bin/python ./tools/app_reg/app_reg.py
    --install
    /opt/cloudera/parcels/SOLR-0.9.0-1.cd4.3.0.pXXX/share/hue/apps/search
```

- ```
c. sed -i 's/\.\/apps/..\./..\./..\./..\./apps/g'
./build/env/lib/python2.X/site-packages/hue.pth
```

where `python2.X` should be the version you are using (for example, `python2.4`).

- ### 3. Start the Hue service.

## The Spark Service

The Spark service is available in two versions: Spark and Spark (Standalone). The previously available Spark service, which runs Spark in standalone mode, has been renamed Spark (Standalone). The Spark (Standalone) service has its own runtime roles: Master and Worker. The current Spark service runs Spark as a YARN application. Both services have a History Server role. In secure clusters, Spark applications can only run on YARN. Cloudera recommends that you use the Spark service.

You can install Spark through the Cloudera Manager Installation wizard using parcels and have the Spark service added and started as part of the Installation wizard. See [Installing Spark](#).

If you elect not to add the Spark service using the Installation wizard, you can use the **Add Service** wizard to create the service. The wizard automatically configures dependent services and the Spark service. See [Adding a Service](#) on page 30 for instructions.

When you upgrade from Cloudera Manager 5.1 or lower to Cloudera 5.2 or higher, Cloudera Manager *does not* migrate an existing Spark service, which runs Spark in standalone mode, to a Spark on YARN service.

### Testing the Spark Service

To test the Spark service, start the Spark shell, `spark-shell`, on one of the hosts. Within the Spark shell, you can run a word count application. For example:

```
val file = sc.textFile("hdfs://namenode:8020/path/to/input")
val counts = file.flatMap(line => line.split(" ")).map(word => (word, 1)).reduceByKey(_ + _)
counts.saveAsTextFile("hdfs://namenode:8020/output")
```

To submit Spark applications to YARN, use the `--master yarn` flag when you start `spark-shell`. To see information about the running Spark shell application, go to Spark History Server UI at `http://spark_history_server:18088` or the [YARN applications](#) page in the Cloudera Manager Admin Console.

If you are running the Spark (Standalone) service, you can see the Spark shell application, and its executors, and logs in the Spark Master UI, by default at `http://spark_master:18080`.

For more information on running Spark applications, see [Running Spark Applications](#).

### Adding the Spark History Server Role

**Required Role:** Cluster Administrator Full Administrator

By default, the Spark (Standalone) service is not created with a History Server. To add the History Server:

1. Go to the Spark service.
2. Click the **Instances** tab.
3. Click the **Add Role Instances** button.
4. Select a host in the column under **History Server**, then click **OK**.
5. Click **Continue**.
6. Check the checkbox next to the History Server role.
7. Select **Actions for Selected** > **Start** and click **Start**.
8. Click **Close** when the action completes.

### The Sqoop 1 Client

The Sqoop 1 client allows you to create a Sqoop 1 [gateway](#) and deploy the client configuration.


#### Installing JDBC Drivers

Sqoop 1 does not ship with third-party JDBC drivers; you must download them separately. For information on downloading and saving the drivers, see [\(CDH 4\) Installing JDBC Drivers](#) and [\(CDH 5\) Installing JDBC Drivers](#). Ensure that you do not save JARs in the CDH parcel directory `/opt/cloudera/parcels/CDH`, because this directory is overwritten when you upgrade CDH.

### Adding the Sqoop 1 Client

**Required Role:** Full Administrator

The Sqoop 1 client packages are installed by the Installation wizard. However, the client configuration is not deployed. To create a Sqoop 1 gateway and deploy the client configuration:

1. On the Home page, click  to the right of the cluster name and select **Add a Service**. A list of service types display. You can add one type of service at a time.
2. Select the **Sqoop 1 Client** service and click **Continue**.
3. Select the radio button next to the services on which the new service should depend and click **Continue**.
4. Customize the assignment of role instances to hosts. The wizard evaluates the hardware configurations of the hosts to determine the best hosts for each role. The wizard assigns all worker roles to the same set of hosts to which the HDFS DataNode role is assigned. These assignments are typically acceptable, but you can reassign them if necessary.

Click a field below a role to display a dialog containing a list of hosts. If you click a field containing multiple hosts, you can also select **All Hosts** to assign the role to all hosts or **Custom** to display the pageable hosts dialog.

The following shortcuts for specifying hostname patterns are supported:

- Range of hostnames (without the domain portion)

| Range Definition        | Matching Hosts                                                                 |
|-------------------------|--------------------------------------------------------------------------------|
| 10.1.1.[1-4]            | 10.1.1.1, 10.1.1.2, 10.1.1.3, 10.1.1.4                                         |
| host[1-3].company.com   | host1.company.com, host2.company.com, host3.company.com                        |
| host[07-10].company.com | host07.company.com, host08.company.com, host09.company.com, host10.company.com |

- IP addresses
- Rack name

Click the **View By Host** button for an overview of the role assignment by hostname ranges.

5. Click **Continue**. The client configuration deployment command runs.
6. Click **Continue** and click **Finish**.

### The Sqoop 2 Service

Cloudera Manager installs the Sqoop 2 service as part of the CDH installation.

You can elect to have the service created and started as part of the Installation wizard. If you elect not to create the service using the Installation wizard, you can use the **Add Service** wizard to perform the installation. The wizard will automatically configure and start the dependent services and the Sqoop 2 service. See [Adding a Service](#) on page 30 for instructions.

### Installing JDBC Drivers

The Sqoop 2 service does not ship with third-party JDBC drivers; you must download them separately. For information on downloading and saving the drivers, see [\(CDH 4\) Configuring Sqoop 2](#) and [\(CDH 5\) Configuring Sqoop 2](#). Ensure that you do not save JARs in the CDH parcel directory `/opt/cloudera/parcels/CDH`, because this directory is overwritten when you upgrade CDH.

### The ZooKeeper Service

**Required Role:** Full Administrator

When adding the ZooKeeper service, the **Add Service** wizard automatically initializes the data directories. If you quit the **Add Service** wizard or it does not finish successfully, you can initialize the directories outside the wizard by doing these steps:

1. Go to the ZooKeeper service.
2. Select **Actions > Initialize**.
3. Click **Initialize** again to confirm.

- **Note:** If the data directories are not initialized, the ZooKeeper servers cannot be started.

In a production environment, you should deploy ZooKeeper as an ensemble with an odd number of servers. As long as a majority of the servers in the ensemble are available, the ZooKeeper service will be available. The minimum recommended ensemble size is three ZooKeeper servers, and Cloudera recommends that each server run on a separate machine. In addition, the ZooKeeper server process should have its own dedicated disk storage if possible.

## Configuring Services to Use the GPL Extras Parcel

After you [install the GPL Extras parcel](#), reconfigure and restart services that need to use LZO functionality. Any service that does not require the use of LZO need not be configured.

### HDFS

1. Go to the HDFS service.
2. Click the **Configuration** tab.
3. Search for the `io.compression.codecs` property.
4. In the **Compression Codecs** property, click in the field, then click the **+** sign to open a new value field.
5. Add the following two codecs:
  - `com.hadoop.compression.lzo.LzoCodec`
  - `com.hadoop.compression.lzo.LzopCodec`
6. Save your configuration changes.
7. Restart HDFS.
8. Redeploy the HDFS client configuration.

### Oozie

1. Go to `/var/lib/oozie` on each Oozie server and even if the LZO JAR is present, symlink the Hadoop LZO JAR:
  - **CDH 5** - `/opt/cloudera/parcels/GPLEXTRAS/lib/hadoop/lib/hadoop-lzo.jar`
  - **CDH 4** - `/opt/cloudera/parcels/HADOOP_LZO/lib/hadoop/lib/hadoop-lzo.jar`
2. Restart Oozie.

### HBase

Restart HBase.

### Impala

Restart Impala.

### Hive

Restart the Hive server.

### Sqoop 2

1. Add the following entries to the **Sqoop Service Environment Advanced Configuration Snippet**:
  - `HADOOP_CLASSPATH=$HADOOP_CLASSPATH:/opt/cloudera/parcels/GPLEXTRAS/lib/hadoop/lib/*`
  - `JAVA_LIBRARY_PATH=$JAVA_LIBRARY_PATH:/opt/cloudera/parcels/GPLEXTRAS/lib/hadoop/lib/native`
2. Restart the Sqoop service.

### Managing Hosts

Cloudera Manager provides a number of features that let you configure and manage the hosts in your clusters.

#### The Status Tab

##### Viewing All Hosts

To display summary information about all the hosts managed by Cloudera Manager, click **Hosts** in the main navigation bar. The All Hosts page displays with a list of all the hosts managed by Cloudera Manager.

The list of hosts shows the overall status of the Cloudera Manager-managed hosts in your cluster. The information provided includes the version of CDH running on the host, the cluster to which the host belongs, and the number of roles running on the host.

- Clicking the ► to the left of the number of roles lists all the role instances running on that host. The balloon annotation that appears when you move the cursor over a link indicates the service instance to which the role belongs.
- Filter the hosts by typing a property value in the Search box or selecting a value from the facets at the left of the page.

##### Disks Overview


Click the **Disks Overview** button to display an overview of the status of all disks in the deployment. The statistics exposed match or build on those in `iostat`, and are shown in a series of histograms that by default cover every physical disk in the system.

Adjust the endpoints of the time line to see the statistics for different time periods. Specify a filter in the box to limit the displayed data. For example, to see the disks for a single rack `rack1`, set the filter to:

`logicalPartition = false and rackId = "rack1"`. Click a histogram to drill down and identify outliers.

##### Viewing the Hosts in a Cluster

Do one of the following:

- Select **Clusters** > **Cluster name** > **General** > **Hosts**.
- In the Home screen, click  **Hosts** in a full form cluster table.

The All Hosts page displays with a list of the hosts filtered by the cluster name.

##### Viewing Individual Hosts

You can view detailed information about an individual host—resources (CPU/memory/storage) used and available, which processes it is running, details about the host agent, and much more—by clicking a host link on the All Hosts page. See [Viewing Host Details](#) on page 97.

#### The Configuration Tab

The **Configuration** tab lets you set properties related to parcels and to resource management, and also monitoring properties for the hosts under management. The configuration settings you make here will affect all your managed hosts. You can also configure properties for individual hosts from the Host Details page (see [Viewing Host Details](#) on page 97) which will override the global properties set here).

To edit the **Default** configuration properties for hosts:

1. Click the **Configuration** tab.

For more information on making configuration changes, see [Modifying Configuration Properties](#) on page 8.



## The Templates Tab

The **Templates** tab lets you create and manage [host templates](#), which provide a way to specify a set of role configurations that should be applied to a host. This greatly simplifies the process of adding new hosts, because it lets you specify the configuration for multiple roles on a host in a single step, and then (optionally) start all those roles.

## The Parcels Tab

In the **Parcels** tab you can download, distribute, and activate available [parcels](#) to your cluster. You can use parcels to add new products to your cluster, or to upgrade products you already have installed.

## Other Host Configuration Tasks

The following topics describe other tasks you can perform concerning the hosts on your cluster.

- [Adding a Host to the Cluster](#) on page 101
- [Specifying Racks for Hosts](#) on page 103
- [Putting a Host into Maintenance Mode](#) on page 106 and [Exiting a Host from Maintenance Mode](#) on page 107
- [Decommissioning and Recommissioning Hosts](#) on page 107
- [Deleting Hosts](#) on page 109
- [Using the Host Inspector](#) on page 100
- [Performing a Rolling Upgrade on a CDH 5 Cluster](#)
- [Re-Running the Cloudera Manager Upgrade Wizard](#)

## Viewing Host Details

You can view detailed information about each host, including:

- Name, IP address, rack ID
- Health status of the host and last time the Cloudera Manager Agent sent a heartbeat to the Cloudera Manager Server
- Number of cores
- System load averages for the past 1, 5, and 15 minutes
- Memory usage
- File system disks, their mount points, and usage
- Health test results for the host
- Charts showing a variety of metrics and health test results over time.
- Role instances running on the host and their health
- CPU, memory, and disk resources used for each role instance



To view detailed host information:

1. Click the **Hosts** tab.
2. Click the name of one of the hosts. The Status page is displayed for the host you selected.
3. Click tabs to access specific categories of information. Each tab provides various categories of information about the host, its services, components, and configuration.

From the status page you can view details about several categories of information.

### Status

The Status page is displayed when a host is initially selected and provides summary information about the status of the selected host. Use this page to gain a general understanding of work being done by the system, the configuration, and health status.

If this host has been decommissioned or is in maintenance mode, you will see the following icon(s) ( ) in the top bar of the page next to the status message.

## Configuring CDH and Managed Services

### Details

This panel provides basic system configuration such as the host's IP address, rack, health status summary, and disk and CPU resources. This information summarizes much of the detailed information provided in other panes on this tab. To view details about the Host agent, click the Host Agent link in the Details section.

### Health Tests

Cloudera Manager monitors a variety of metrics that are used to indicate whether a host is functioning as expected. The Health Tests panel shows health test results in an expandable/collapsible list, typically with the specific metrics that the test returned. (You can Expand All or Collapse All from the links at the upper right of the Health Tests panel).

- The color of the text (and the background color of the field) for a health test result indicates the status of the results. The tests are sorted by their health status – Good, Concerning, Bad, or Disabled. The list of entries for good and disabled health tests are collapsed by default; however, Bad or Concerning results are shown expanded.
- The text of a health test also acts as a link to further information about the test. Clicking the text will pop up a window with further information, such as the meaning of the test and its possible results, suggestions for actions you can take or how to make configuration changes related to the test. The help text for a health test also provides a link to the relevant monitoring configuration section for the service. See [Configuring Monitoring Settings](#) for more information.

### Health History

The Health History provides a record of state transitions of the health tests for the host.

- Click the arrow symbol at the left to view the description of the health test state change.
- Click the **View** link to open a new page that shows the state of the host at the time of the transition. In this view some of the status settings are greyed out, as they reflect a time in the past, not the current status.

### File Systems

The File systems panel provides information about disks, their mount points and usage. Use this information to determine if additional disk space is required.

### Roles

Use the Roles panel to see the role instances running on the selected host, as well as each instance's status and health. Hosts are configured with one or more role instances, each of which corresponds to a service. The role indicates which daemon runs on the host. Some examples of roles include the NameNode, Secondary NameNode, Balancer, JobTrackers, DataNodes, RegionServers and so on. Typically a host will run multiple roles in support of the various services running in the cluster.

Clicking the role name takes you to the role instance's status page.

You can delete a role from the host from the Instances tab of the Service page for the parent service of the role. You can add a role to a host in the same way. See [Role Instances](#) on page 40.

### Charts

Charts are shown for each host instance in your cluster.

See [Viewing Charts for Cluster, Service, Role, and Host Instances](#) for detailed information on the charts that are presented, and the ability to search and display metrics of your choice.

### Processes

The Processes page provides information about each of the processes that are currently running on this host. Use this page to access management web UIs, check process status, and access log information.

- **Note:** The Processes page may display exited startup processes. Such processes are cleaned up within a day.

The Processes tab includes a variety of categories of information.

- **Service** - The name of the service. Clicking the service name takes you to the service status page. Using the triangle to the right of the service name, you can directly access the tabs on the role page (such as the Instances, Commands, Configuration, Audits, or Charts Library tabs).
- **Instance** - The role instance on this host that is associated with the service. Clicking the role name takes you to the role instance's status page. Using the triangle to the right of the role name, you can directly access the tabs on the role page (such as the Processes, Commands, Configuration, Audits, or Charts Library tabs) as well as the status page for the parent service of the role.
- **Name** - The process name.
- **Links** - Links to management interfaces for this role instance on this system. These is not available in all cases.
- **Status** - The current status for the process. Statuses include stopped, starting, running, and paused.
- **PID** - The unique process identifier.
- **Uptime** - The length of time this process has been running.
- **Full log file** - A link to the full log (a file external to Cloudera Manager) for this host log entries for this host.
- **Stderr** - A link to the stderr log (a file external to Cloudera Manager) for this host.
- **Stdout** - A link to the stdout log (a file external to Cloudera Manager) for this host.

## Resources

The Resources page provides information about the resources (CPU, memory, disk, and ports) used by every service and role instance running on the selected host.

Each entry on this page lists:

- The service name
- The name of the particular instance of this service
- A brief description of the resource
- The amount of the resource being consumed or the settings for the resource

The resource information provided depends on the type of resource:

- **CPU** - An approximate percentage of the CPU resource consumed.
- **Memory** - The number of bytes consumed.
- **Disk** - The disk location where this service stores information.
- **Ports** - The port number being used by the service to establish network connections.

## Commands

The Commands page shows you running or recent commands for the host you are viewing. See [Viewing Running and Recent Commands](#) for more information.

## Configuration

[Required Role:](#) **Full Administrator**

The Configuration page for a host lets you set monitoring properties for the selected host. In addition, for parcel upgrades, you can blacklist specific products - specify products that should not be distributed or activated on the host.

To modify the monitoring properties for the selected host:

1. Click the **Configuration** tab.
2. Click the **Monitoring** category.

## Configuring CDH and Managed Services

- Under **Thresholds** you can configure the thresholds for monitoring the free space in the Agent Log and Agent Process Directories for all your hosts. You can set these thresholds as either or both a percentage and an absolute value (in bytes).
- Under **Other** you can set health check thresholds for a variety of conditions related to memory usage and other properties. Here is where you can enable Alerting for health check events for all your managed hosts.

The monitoring settings you make on this page will override the global host monitoring settings from the Configuration tab of the Hosts page.

For more information, see [Modifying Configuration Properties](#) on page 8.

### Components

The Components page lists every component installed on this host. This may include components that have been installed but have not been added as a service (such as YARN, Flume, or Impala).

This includes the following information:

- **Component** - The name of the component.
- **Version** - The version of CDH from which each component came.
- **Component Version** - The detailed version number for each component.

### Audits

The Audits page lets you filter for audit events related to this host. See [Audit Events](#) for more information.

### Charts Library

The Charts Library page for a host instance provides charts for all metrics kept for that host instance, organized by category. Each category is collapsible/expandable. See [Viewing Charts for Cluster, Service, Role, and Host Instances](#) for more information.

## Using the Host Inspector

[Required Role:](#) **Full Administrator**

You can use the host inspector to gather information about hosts that Cloudera Manager is currently managing. You can review this information to better understand system status and troubleshoot any existing issues. For example, you might use this information to investigate potential DNS misconfiguration.

The inspector runs tests to gather information for functional areas including:

- Networking
- System time
- User and group configuration
- HDFS settings
- Component versions

Common cases in which this information is useful include:

- Installing components
- Upgrading components
- Adding hosts to a cluster
- Removing hosts from a cluster

### Running the Host Inspector

1. Click the **Hosts** tab.
2. Click **Host Inspector**. Cloudera Manager begins several tasks to inspect the managed hosts.
3. After the inspection completes, click **Download Result Data** or **Show Inspector Results** to review the results.

The results of the inspection displays a list of all the validations and their results, and a summary of all the components installed on your managed hosts.

If the validation process finds problems, the **Validations** section will indicate the problem. In some cases the message may indicate actions you can take to resolve the problem. If an issue exists on multiple hosts, you may be able to view the list of occurrences by clicking a small triangle that appears at the end of the message.

The **Version Summary** section shows all the components that are available from Cloudera, their versions (if known) and the CDH distribution to which they belong (CDH 4 or CDH 5).

If you are running multiple clusters with both CDH 4 and CDH 5, the lists will be organized by distribution (CDH 4 or CDH 5). The hosts running that version are shown at the top of each list.

### Viewing Past Host Inspector Results

You can view the results of a past host inspection by looking for the Host Inspector command using the **Recent Commands** feature.

1. Click the Running Commands indicator (1) just to the left of the Search box at the right hand side of the navigation bar.
2. Click the **Recent Commands** button.
3. If the command is too far in the past, you can use the Time Range Selector to move the time range back to cover the time period you want.
4. When you find the Host Inspector command, click its name to display its subcommands.
5. Click the **Show Inspector Results** button to view the report.

See [Viewing Running and Recent Commands](#) for more information about viewing past command activity.

### Adding a Host to the Cluster

**Required Role:** Full Administrator

You can add one or more hosts to your Hadoop cluster using the Add Hosts wizard, which will install the Oracle JDK, CDH, Impala (optional) and the Cloudera Manager Agent packages. After these packages are installed and the Cloudera Manager Agent is started, the Agent will connect to the Cloudera Manager Server and you will then be able to use the Cloudera Manager Admin Console to manage and monitor CDH on the new host.

The Add Hosts wizard does not create roles on the new host; once you have successfully added the host(s) you can either add roles, one service at a time, or apply a host template, which can define role configurations for multiple roles.

#### ■ Important:

- All hosts in a single cluster must be running the same version of CDH, for example CDH 4.5 or CDH 5.0.
- When you install the new hosts on your system, you must install the same version of CDH to enable the new host to work with the other hosts in the cluster. The installation wizard lets you select the version of CDH you want to install, and you can choose a custom repository to ensure that the version you install matches the version on your other hosts.
- If you are managing multiple clusters, be sure to select the version of CDH that matches the version in use on the cluster where you plan to add the new hosts.

### Using the Add Hosts Wizard to Add Hosts

You can use the Add Hosts wizard to install CDH, Impala, and the Cloudera Manager Agent on a host.

### Disable TLS Encryption or Authentication

If you have enabled TLS encryption or authentication for the Cloudera Manager Agents, you must disable both of them before starting the Add Hosts wizard. Otherwise, skip to the next step.

- **Important:** This step leaves the existing hosts in an unmanageable state; they are still configured to use TLS, and so can't communicate with the Cloudera Manager Server.

1. From the **Administration** tab, select **Settings**.
2. Select the **Security** category.
3. Disable all levels of TLS that are currently enabled by deselecting the following options: **Use TLS Encryption for Agents**, and **Use TLS Authentication of Agents to Server**.
4. Click **Save Changes** to save the settings.
5. Restart the Cloudera Management Server to have these changes take effect.

### Using the Add Hosts Wizard

1. Click the **Hosts** tab.
2. Click the **Add New Hosts** button.
3. Follow the instructions in the wizard to install the Oracle JDK, CDH, managed service, and Cloudera Manager Agent packages or parcels and start the Agent.
4. In the **Specify hosts for your CDH Cluster installation** page, you can search for new hosts to add under the **New Hosts** tab. However, if you have hosts that are already known to Cloudera Manager but have no roles assigned, (for example, a host that was previously in your cluster but was then removed) these will appear under the **Currently Managed Hosts** tab.
5. You will have an opportunity to add (and start) role instances to your newly-added hosts using a host template.
  - a. You can select an existing host template, or create a new one.
  - b. To create a new host template, click the **+ Create...** button. This will open the **Create New Host Template** pop-up. See [Host Templates](#) on page 103 for details on how you select the role groups that define the roles that should run on a host. When you have created the template, it will appear in the list of host templates from which you can choose.
  - c. Select the host template you want to use.
  - d. By default Cloudera Manager will automatically start the roles specified in the host template on your newly added hosts. To prevent this, uncheck the option to start the newly-created roles.
6. When the wizard is finished, you can verify the Agent is connecting properly with the Cloudera Manager Server by clicking the **Hosts** tab and checking the health status for the new host. If the Health Status is **Good** and the value for the Last Heartbeat is recent, then the Agent is connecting properly with the Cloudera Manager Server.

If you did not specify a host template during the Add Hosts wizard, then no roles will be present on your new hosts until you add them. You can do this by adding individual roles under the **Instances** tab for a specific service, or by using a host template. See [Role Instances](#) on page 40 for information about adding roles for a specific service. See [Host Templates](#) on page 103 to create a host template that specifies a set of roles (from different services) that should run on a host.

### Enable TLS Encryption or Authentication

If you previously enabled TLS security on your cluster, you must re-enable the TLS options on the **Administration** page and also configure TLS on each new host after using the Add Hosts wizard. Otherwise, you can ignore this step.

### Adding a Host by Installing the Packages Using Your Own Method

If you used a different mechanism to install the Oracle JDK, CDH, Cloudera Manager Agent packages, you can use that same mechanism to install the Oracle JDK, CDH, Cloudera Manager Agent packages and then start the Cloudera Manager Agent.

1. Install the Oracle JDK, CDH, and Cloudera Manager Agent packages using your own method. For instructions on installing these packages, see [Installation Path B - Manual Installation Using Cloudera Manager Packages](#).

2. After installation is complete, start the Cloudera Manager Agent. For instructions, see [Starting, Stopping, and Restarting Cloudera Manager Agents](#) on page 296.
3. After the Agent is started, you can verify the Agent is connecting properly with the Cloudera Manager Server by clicking the **Hosts** tab and checking the health status for the new host. If the Health Status is **Good** and the value for the Last Heartbeat is recent, then the Agent is connecting properly with the Cloudera Manager Server.
4. If you have enabled TLS security on your cluster, you must enable and configure TLS on each new host. Otherwise, ignore this step.
5. Enable and configure TLS on each new host by specifying 1 for the `use_tls` property in the `/etc/cloudera-scm-agent/config.ini` configuration file.
6. Configure the same level(s) of TLS security on the new hosts by following the instructions in [Configuring TLS Security for Cloudera Manager](#).

## Specifying Racks for Hosts

Required Role: **Cluster Administrator** **Full Administrator**

To get maximum performance, it is important to configure CDH so that it knows the topology of your network. Network locations such as hosts and racks are represented in a tree, which reflects the network “distance” between locations. HDFS will use the network location to be able to place block replicas more intelligently to trade off performance and resilience. When placing jobs on hosts, CDH will prefer within-rack transfers (where there is more bandwidth available) to off-rack transfers; the MapReduce and YARN schedulers use network location to determine where the closest replica is as input to a map task. These computations are performed with the assistance of rack awareness scripts.

Cloudera Manager includes internal rack awareness scripts, but you must specify the racks where the hosts in your cluster are located. If your cluster contains more than 10 hosts, Cloudera recommends that you specify the rack for each host. HDFS, MapReduce, and YARN will automatically use the racks you specify.

Cloudera Manager supports nested rack specifications. For example, you could specify the rack `/rack3`, or `/group5/rack3` to indicate the third rack in the fifth group. All hosts in a cluster must have the *same number* of path components in their rack specifications.

To specify racks for hosts:

1. Click the **Hosts** tab.
2. Check the checkboxes next to the host(s) for a particular rack, such as all hosts for `/rack123`.
3. Click **Actions for Selected (n) > Assign Rack**, where *n* is the number of selected hosts.
4. Enter a rack name or ID that starts with a slash `/`, such as `/rack123` or `/aisle1/rack123`, and then click **Confirm**.
5. Optionally [restart affected services](#). Rack assignments are not automatically updated for running services.

## Host Templates

Required Role: **Full Administrator**

Host templates let you designate a set of role groups that can be applied in a single operation to a host or a set of hosts. This significantly simplifies the process of configuring new hosts when you need to expand your cluster. Host templates are supported for both CDH 4 and CDH 5 cluster hosts.

- **Important:** A host template can only be applied on a host with a version of CDH that matches the CDH version running on the cluster to which the host template belongs.

You can create and manage host templates under the Templates tab from the Hosts page.

1. Click the **Hosts** tab on the main Cloudera Manager navigation bar.
2. Click the **Templates** tab on the Hosts page.

Templates are not required; Cloudera Manager assigns roles and role groups to the hosts of your cluster when you perform the initial cluster installation. However, if you want to add new hosts to your cluster, a host template can make this much easier.

If there are existing host templates, they are listed on the page, along with links to each role group included in the template.

If you are managing multiple clusters, you must create separate host templates for each cluster, as the templates specify role configurations specific to the roles in a single cluster. Existing host templates are listed under the cluster to which they apply.

- You can click a role group name to be taken to the Edit configuration page for that role group, where you can modify the role group settings.
- From the **Actions** menu associated with the template you can edit the template, clone it, or delete it.

### Creating a Host Template

1. From the **Templates** tab, click **Click here**
2. In the **Create New Host Template** pop-up window that appears:
  - Type a name for the template.
  - For each role, select the appropriate role group. There may be multiple role groups for a given role type — you want to select the one with the configuration that meets your needs.
3. Click **Create** to create the host template.

### Editing a Host Template

1. From the **Hosts** tab, click the **Templates** tab.
2. Pull down the **Actions** menu for the template you want to modify, and click **Edit**. This put you into the **Edit Host Template** pop-up window. This works exactly like the **Create New Host Template** window — you can modify the template name or any of the role group selections.
3. Click **OK** when you have finished.

### Applying a Host Template to a Host

You can use a host template to apply configurations for multiple roles in a single operation.

You can apply a template to a host that has no roles on it, or that has roles from the same services as those included in the host template. New roles specified in the template that do not already exist on the host will be added. A role on the host that is already a member of the role group specified in the template will be left unchanged. If a role on the host matches a role in the template, but is a member of a different role group, it will be moved to the role group specified by the template.

For example, suppose you have two role groups for a DataNode (DataNode Default Group and DataNode (1)). The host has a DataNode role that belongs to DataNode Default Group. If you apply a host template that specifies the DataNode (1) group, the role on the host will be moved from DataNode Default Group to DataNode (1).

However, if you have two instances of a service, such as MapReduce (for example, *mr1* and *mr2*) and the host has a TaskTracker role from service *mr2*, you cannot apply a TaskTracker role from service *mr1*.

A host may have no roles on it if you have just added the host to your cluster, or if you decommissioned a managed host and removed its existing roles.

Also, the host must have the same version of CDH installed as is running on the cluster whose host templates you are applying.

If a host belongs to a different cluster than the one for which you created the host template, you can apply the host template if the "foreign" host either has no roles on it, or has only management roles on it. When you apply the host template, the host will then become a member of the cluster whose host template you applied. The following instructions assume you have already created the appropriate host template.

1. Go to the **Hosts** page, **Status** tab.
2. Select the host(s) to which you want to apply your host template.



3. From the **Actions for Selected** menu, select **Apply Host Template**.
4. In the pop-up window that appears, select the host template you want to apply.
5. Optionally you can have Cloudera Manager start the roles created per the host template – check the box to enable this.
6. Click **Confirm** to initiate the action.

## Maintenance Mode

**Required Role:** Configurator Cluster Administrator Full Administrator

Maintenance mode allows you to suppress alerts for a host, service, role, or an entire cluster. This can be useful when you need to take actions in your cluster (make configuration changes and restart various elements) and do not want to see the alerts that will be generated due to those actions.



Putting an entity into maintenance mode does not prevent events from being logged; it only suppresses the alerts that those events would otherwise generate. You can see a history of all the events that were recorded for entities during the period that those entities were in maintenance mode.

### Explicit and Effective Maintenance Mode

When you enter maintenance mode on an entity (cluster, service, or host) that has subordinate entities (for example, the roles for a service) the subordinate entities are also put into maintenance mode. These are considered to be in **effective maintenance mode**, as they have inherited the setting from the higher-level entity.

For example:


- If you set the HBase service into maintenance mode, then its roles (HBase Master and all RegionServers) are put into effective maintenance mode.
- If you set a host into maintenance mode, then any roles running on that host are put into effective maintenance mode.

Entities that have been explicitly put into maintenance mode show the icon . Entities that have entered effective maintenance mode as a result of inheritance from a higher-level entity show the icon .

When an entity (role, host or service) is in effective maintenance mode, it can only be removed from maintenance mode when the higher-level entity exits maintenance mode. For example, if you put a service into maintenance mode, then the roles associated with that service will be entered into effective maintenance mode, and will remain in effective maintenance mode until the service exits maintenance mode. You cannot remove them from maintenance mode individually.

On the other hand, an entity that is in effective maintenance mode can be put into explicit maintenance mode. In this case, the entity will remain in maintenance mode even when the higher-level entity exits maintenance mode. For example, suppose you put a host into maintenance mode, (which puts all the roles on that host into effective maintenance mode). You then select one of the roles on that host and put it explicitly into maintenance mode. When you have the host exit maintenance mode, that one role will remain in maintenance mode. You will need to select it individually and specifically have it exit maintenance mode.

### Viewing Maintenance Mode Status

You can view the status of Maintenance Mode in your cluster by clicking  to the right of the cluster name and selecting **View Maintenance Mode Status**.

### Entering Maintenance Mode



You can enable maintenance mode for a cluster, service, role, or host.

### Putting a Cluster into Maintenance Mode


1. Click  to the right of the cluster name and select **Enter Maintenance Mode**.



## Configuring CDH and Managed Services

2. Confirm that you want to do this.

The cluster is put into explicit maintenance mode, as indicated by the  icon. All services and roles in the cluster are entered into effective maintenance mode, as indicated by the  icon.

### Putting a Service into Maintenance Mode

1. Click  to the right of the service name and select **Enter Maintenance Mode**.
2. Confirm that you want to do this.

The service is put into explicit maintenance mode, as indicated by the  icon. All roles for the service are entered into effective maintenance mode, as indicated by the  icon.

### Putting Roles into Maintenance Mode

1. Go to the service page that includes the role.
2. Go to the **Instances** tab.
3. Select the role(s) you want to put into maintenance mode.
4. From the **Actions for Selected** menu, select **Enter Maintenance Mode**.
5. Confirm that you want to do this.

The roles will be put in explicit maintenance mode. If the roles were already in effective maintenance mode (because its service or host was put into maintenance mode) the roles will now be in explicit maintenance mode. This means that they will not exit maintenance mode automatically if their host or service exits maintenance mode; they must be explicitly removed from maintenance mode.

### Putting a Host into Maintenance Mode


1. Go to the **Hosts** page.
2. Select the host(s) you want to put into maintenance mode.
3. From the **Actions for Selected** menu, select **Enter Maintenance Mode**.
4. Confirm that you want to do this.

The confirmation pop-up lists the role instances that will be put into effective maintenance mode when the host goes into maintenance mode.


### Exiting Maintenance Mode

When you exit maintenance mode, the maintenance mode icons are removed and alert notification resumes.

### Exiting a Cluster from Maintenance Mode

1. Click  to the right of the cluster name and select **Exit Maintenance Mode**.
2. Confirm that you want to do this.

### Exiting a Service from Maintenance Mode

1. Click  to the right of the service name and select **Exit Maintenance Mode**.
2. Confirm that you want to do this.

### Exiting Roles from Maintenance Mode

1. Go to the services page that includes the role.
2. Go to the **Instances** tab.

3. Select the role(s) you want to exit from maintenance mode.
4. From the **Actions for Selected** menu, select **Exit Maintenance Mode**.
5. Confirm that you want to do this.

#### Exiting a Host from Maintenance Mode

1. Go to the **Hosts** page.
2. Select the host(s) you want to put into maintenance mode.
3. From the **Actions for Selected** menu, select **Exit Maintenance Mode**.
4. Confirm that you want to do this.

The confirmation pop-up lists the role instances that will be removed from effective maintenance mode when the host exits maintenance mode.

### Decommissioning and Recommissioning Hosts

Decommissioning a host decommissions and stops all roles on the host without having to go to each service and individually decommission the roles. Decommissioning applies to only to HDFS DataNode, MapReduce TaskTracker, YARN NodeManager, and HBase RegionServer roles. If the host has other roles running on it, those roles are stopped.

Once all roles on the host have been decommissioned and stopped, the host can be removed from service. You can decommission multiple hosts in parallel.

#### Decommissioning Hosts

**Required Role:** Limited Operator Operator Configurator Cluster Administrator Full Administrator



You cannot decommission a DataNode or a host with a DataNode if the number of DataNodes equals the replication factor (which by default is three) of any file stored in HDFS. For example, if the replication factor of any file is three, and you have three DataNodes, you cannot decommission a DataNode or a host with a DataNode.

To decommission hosts:

1. If the host has a DataNode, perform the steps in [Tuning HDFS Prior to Decommissioning DataNodes](#) on page 108.
2. Click the **Hosts** tab.
3. Check the checkboxes next to one or more hosts.
4. Select **Actions for Selected** > **Decommission**.

A confirmation pop-up informs you of the roles that will be decommissioned or stopped on the hosts you have selected. To proceed with the decommissioning, click **Confirm**.

A Command Details window appears that will show each stop or decommission command as it is run, service by service. You can click one of the decommission links to see the subcommands that are run for decommissioning a given role. Depending on the role, the steps may include adding the host to an "exclusions list" and refreshing the NameNode, JobTracker, or NodeManager, stopping the Balancer (if it is running), and moving data blocks or regions. Roles that do not have specific decommission actions are stopped.

While decommissioning is in progress, the host displays the  icon. Once all roles have been decommissioned or stopped, the host displays the  icon. If one host in a cluster has been decommissioned, the DECOMMISSIONED facet displays in the Filters on the Hosts page and you can filter the hosts according to their decommission status.

You cannot start roles on a decommissioned host.

# Configuring CDH and Managed Services

## Tuning HDFS Prior to Decommissioning DataNodes

**Required Role:** Configurator Cluster Administrator Full Administrator

When a DataNode is decommissioned, the NameNode ensures that every block from the DataNode will still be available across the cluster as dictated by the replication factor. This procedure involves copying blocks off the DataNode in small batches. In cases where a DataNode has thousands of blocks, decommissioning can take several hours. Before decommissioning hosts with DataNodes, you should first tune HDFS:


1. Raise the heap size of the DataNodes. DataNodes should be configured with at least 4 GB heap size to allow for the increase in iterations and max streams.
  - a. Go to the HDFS service page.
  - b. Click the **Configuration** tab.
  - c. Under each DataNode role group (DataNode Default Group and any additional DataNode role groups) go to the **Resource Management** category, and set the **Java Heap Size of DataNode in Bytes** property as recommended.
  - d. Click **Save Changes** to commit the changes.
2. Set the DataNode balancing bandwidth:
  - a. Expand the **DataNode Default Group > Performance** category.
  - b. Configure the **DataNode Balancing Bandwidth** property to the bandwidth you have on your disks and network.
  - c. Click **Save Changes** to commit the changes.
3. Increase the replication work multiplier per iteration to a larger number (the default is 2, however 10 is recommended):
  - a. Expand the **NameNode Default Group > Advanced** category.
  - b. Configure the **Replication Work Multiplier Per Iteration** property to a value such as 10.
  - c. Click **Save Changes** to commit the changes.
4. Increase the replication maximum threads and maximum replication thread hard limits:
  - a. Expand the **NameNode Default Group > Advanced** category.
  - b. Configure the **Maximum number of replication threads on a Datanode** and **Hard limit on the number of replication threads on a Datanode** properties to 50 and 100 respectively.
  - c. Click **Save Changes** to commit the changes.
5. Restart the HDFS service.

## Recommissioning Hosts

**Required Role:** Operator Configurator Cluster Administrator Full Administrator

Only hosts that are decommissioned using Cloudera Manager can be recommissioned.

1. Click the **Hosts** tab.
2. Select one or more hosts to recommission.
3. Select **Actions for Selected > Recommission**.

The  icon is removed from the host and from the roles that reside on the host. However, the roles themselves are not restarted.

## Restarting All The Roles on a Recommissioned Host

**Required Role:** Operator Configurator Cluster Administrator Full Administrator

1. Click the **Hosts** tab.

2. Select one or more hosts on which to start recommissioned roles.
3. Select **Actions for Selected** > **Start All Roles**.

## Deleting Hosts

Required Role: **Full Administrator**

There are two ways to remove a host from a cluster:

- Stop Cloudera Manager from managing a host entirely, and the Hadoop daemons on the host.
- Remove a host from a cluster, but leave it available to be added to a different cluster managed by Cloudera Manager.

### Delete a Host From Being Managed by Cloudera Manager

1. In the Cloudera Manager Admin Console, click the **Hosts** tab.
2. Select the hosts to delete.
3. Select **Actions for Selected** > **Decommission** to ensure that all roles on the host have been stopped. For further details, see [Decommissioning and Recommissioning Hosts](#) on page 107.
4. Stop the Agent on the host. For instructions, see [Starting, Stopping, and Restarting Cloudera Manager Agents](#) on page 296.
5. In the Cloudera Manager Admin Console, click the **Hosts** tab and select the host you want to delete.
6. Select **Actions for Selected** > **Delete**.

### Removing a Host From a Cluster But Leave It Available To Cloudera Manager

You can remove a host from a cluster, but leaving it managed by Cloudera Manager so it can be added a different cluster being managed by the same Cloudera Manager server. This command stops all running roles and deletes, them and then removes the host from the cluster. It preserves the role data directories. You can choose to have Cloudera Manager management roles (such as the Events Server, Activity Monitor and so on), remain. The Cloudera Manager Agent should continue to run on the host.

1. In the Cloudera Manager Admin Console, click the **Hosts** tab.
2. Select the hosts to delete.
3. Select **Actions for Selected** > **Remove From Cluster**.
4. In the pop-up that appears, make your selections for decommissioning role (default is to do this) and whether to skip removing the management service roles (default is to leave them). Click **Confirm** to proceed with removing the selected hosts.

## Configuring and Using CDH from the Command Line

The following sections provide instructions and information on configuring core Hadoop.

For installation and upgrade instructions, see the [Cloudera Installation and Upgrade](#) guide, which also contains initial deployment and configuration instructions for core Hadoop and the CDH components, including:

- Cluster configuration and maintenance:
  - [Ports Used by Components of CDH 5](#)
  - [Configuring Network Names](#)
  - [Deploying CDH 5 on a Cluster](#)
  - [Starting CDH Services](#) on page 110
  - [Stopping Services](#) on page 115
- [Avro Usage](#)
- [Flume configuration](#)
- HBase configuration:

## Configuring CDH and Managed Services

- [Configuration Settings for HBase](#)
- [HBase Replication](#) on page 269
- [HBase Snapshots](#)
- [HCatalog configuration](#)
- [Impala configuration](#)
- Hive configuration:
  - [Configuring the Hive Metastore](#)
  - [Configuring HiveServer2](#)
  - [Configuring the Metastore to use HDFS High Availability](#)
- [HttpFS configuration](#)
- Hue: [Configuring CDH Components for Hue](#)
- Oozie configuration:
  - [Configuring Oozie](#)
  - [Configuring Oozie Failover \(hot/cold\)](#)
- Parquet: [Using the Parquet File Format with Impala, Hive, Pig, HBase, and MapReduce](#)
- Snappy:
  - [Using Snappy for MapReduce Compression](#)
  - [Using Snappy for Pig Compression](#)
  - [Using Snappy for Hive Compression](#)
  - [Using Snappy Compression in Sqoop 1 and Sqoop 2 Imports](#)
  - [Using Snappy Compression with HBase](#)
- Spark configuration:
  - [Configuring and Running Spark \(Standalone Mode\)](#)
  - [Running Spark Applications](#)
  - [Running Crunch with Spark](#)
- Sqoop configuration:
  - [Setting HADOOP\\_MAPRED\\_HOME](#) for Sqoop
  - [Configuring Sqoop 2](#)
- ZooKeeper: [Maintaining a ZooKeeper Server](#)

## Starting CDH Services

You need to start and stop services in the right order to make sure everything starts or stops cleanly.

- **Note:** The Oracle JDK is required for all Hadoop components.

START services in this order:

| Order | Service   | Comments                                                                                                                                                       | For instructions and more information                                                                                                                                                                                      |
|-------|-----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1     | ZooKeeper | Cloudera recommends starting ZooKeeper before starting HDFS; this is a <b>requirement</b> in a <a href="#">high-availability (HA)</a> deployment. In any case, | <a href="#">Installing the ZooKeeper Server Package and Starting ZooKeeper on a Single Server</a> ; <a href="#">Installing ZooKeeper in a Production Environment</a> ; <a href="#">Deploying HDFS High Availability</a> on |

| Order | Service   | Comments                                                                                                                                            | For instructions and more information                                                                                                |
|-------|-----------|-----------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------|
|       |           | always start ZooKeeper before HBase.                                                                                                                | page 224; <a href="#">Configuring High Availability for the JobTracker (MRv1)</a>                                                    |
| 2     | HDFS      | Start HDFS before all other services except ZooKeeper. If you are using HA, see the <a href="#">CDH 5 High Availability Guide</a> for instructions. | <a href="#">Deploying HDFS on a Cluster</a> ; <a href="#">HDFS High Availability</a> on page 211                                     |
| 3     | HttpFS    |                                                                                                                                                     | <a href="#">HttpFS Installation</a>                                                                                                  |
| 4a    | MRv1      | Start MapReduce before Hive or Oozie. Do not start MRv1 if YARN is running.                                                                         | <a href="#">Deploying MapReduce v1 (MRv1) on a Cluster</a> ; <a href="#">Configuring High Availability for the JobTracker (MRv1)</a> |
| 4b    | YARN      | Start YARN before Hive or Oozie. Do not start YARN if MRv1 is running.                                                                              | <a href="#">Deploying MapReduce v2 (YARN) on a Cluster</a>                                                                           |
| 5     | HBase     |                                                                                                                                                     | <a href="#">Starting the HBase Master</a> ; <a href="#">Deploying HBase in a Distributed Cluster</a>                                 |
| 6     | Hive      | Start the Hive metastore before starting HiveServer2 and the Hive console.                                                                          | <a href="#">Installing Hive</a>                                                                                                      |
| 7     | Oozie     |                                                                                                                                                     | <a href="#">Starting the Oozie Server</a>                                                                                            |
| 8     | Flume 1.x |                                                                                                                                                     | <a href="#">Running Flume</a>                                                                                                        |
| 9     | Sqoop     |                                                                                                                                                     | <a href="#">Sqoop Installation</a> and <a href="#">Sqoop 2 Installation</a>                                                          |
| 10    | Hue       |                                                                                                                                                     | <a href="#">Hue Installation</a>                                                                                                     |

## Configuring init to Start Hadoop System Services

### ■ Important:

- If you use Cloudera Manager, do not use these command-line instructions.
- This information applies specifically to CDH 5.3.x . If you use an earlier version of CDH, see the documentation for that version located at [Cloudera Documentation](#).

`init(8)` starts some daemons when the system is booted. Depending on the distribution, `init` executes scripts from either the `/etc/init.d` directory or the `/etc/rc2.d` directory. The CDH packages link the files in `init.d` and `rc2.d` so that modifying one set of files automatically updates the other.

To start system services at boot time and on restarts, enable their `init` scripts on the systems on which the services will run, using the appropriate tool:

## Configuring CDH and Managed Services

- `chkconfig` is included in the RHEL and CentOS distributions. Debian and Ubuntu users can install the `chkconfig` package.
- `update-rc.d` is included in the Debian and Ubuntu distributions.

### Configuring init to Start Core Hadoop System Services in an MRv1 Cluster

- **Important:**  
Cloudera does not support running MRv1 and YARN daemons on the same nodes at the same time; it will degrade performance and may result in an unstable cluster deployment.

The `chkconfig` commands to use are:

```
$ sudo chkconfig hadoop-hdfs-namenode on
```

The `update-rc.d` commands to use on Ubuntu and Debian systems are:

| Where                               | Command                                                                   |
|-------------------------------------|---------------------------------------------------------------------------|
| On the NameNode                     | <pre>\$ sudo update-rc.d hadoop-hdfs-namenode defaults</pre>              |
| On the JobTracker                   | <pre>\$ sudo update-rc.d hadoop-0.20-mapreduce-jobtracker defaults</pre>  |
| On the Secondary NameNode (if used) | <pre>\$ sudo update-rc.d hadoop-hdfs-secondarynamenode defaults</pre>     |
| On each TaskTracker                 | <pre>\$ sudo update-rc.d hadoop-0.20-mapreduce-tasktracker defaults</pre> |
| On each DataNode                    | <pre>\$ sudo update-rc.d hadoop-hdfs-datanode defaults</pre>              |

### Configuring init to Start Core Hadoop System Services in a YARN Cluster

- **Important:**  
Do not run MRv1 and YARN on the same set of nodes at the same time. This is not recommended; it degrades your performance and may result in an unstable MapReduce cluster deployment.

The `chkconfig` commands to use are:



| Where                               | Command                                                              |
|-------------------------------------|----------------------------------------------------------------------|
| On the NameNode                     | <code>\$ sudo chkconfig hadoop-hdfs-namenode on</code>               |
| On the ResourceManager              | <code>\$ sudo chkconfig<br/>hadoop-yarn-resourcemanager on</code>    |
| On the Secondary NameNode (if used) | <code>\$ sudo chkconfig<br/>hadoop-hdfs-secondarynamenode on</code>  |
| On each NodeManager                 | <code>\$ sudo chkconfig hadoop-yarn-nodemanager<br/>on</code>        |
| On each DataNode                    | <code>\$ sudo chkconfig hadoop-hdfs-datanode on</code>               |
| On the MapReduce JobHistory node    | <code>\$ sudo chkconfig<br/>hadoop-mapreduce-historyserver on</code> |

The `update-rc.d` commands to use on Ubuntu and Debian systems are:

| Where                               | Command                                                                      |
|-------------------------------------|------------------------------------------------------------------------------|
| On the NameNode                     | <code>\$ sudo update-rc.d hadoop-hdfs-namenode<br/>defaults</code>           |
| On the ResourceManager              | <code>\$ sudo update-rc.d<br/>hadoop-yarn-resourcemanager defaults</code>    |
| On the Secondary NameNode (if used) | <code>\$ sudo update-rc.d<br/>hadoop-hdfs-secondarynamenode defaults</code>  |
| On each NodeManager                 | <code>\$ sudo update-rc.d hadoop-yarn-nodemanager<br/>defaults</code>        |
| On each DataNode                    | <code>\$ sudo update-rc.d hadoop-hdfs-datanode<br/>defaults</code>           |
| On the MapReduce JobHistory node    | <code>\$ sudo update-rc.d<br/>hadoop-mapreduce-historyserver defaults</code> |

### Configuring init to Start Non-core Hadoop System Services

Non-core Hadoop daemons can also be configured to start at `init` time using the `chkconfig` or `update-rc.d` command.

The `chkconfig` commands are:

| Component      | Server                | Command                                              |
|----------------|-----------------------|------------------------------------------------------|
| Hue            | Hue server            | <code>\$ sudo chkconfig hue on</code>                |
| Oozie          | Oozie server          | <code>\$ sudo chkconfig oozie on</code>              |
| HBase          | HBase master          | <code>\$ sudo chkconfig hbase-master on</code>       |
|                | On each HBase slave   | <code>\$ sudo chkconfig hbase-regionserver on</code> |
| Hive Metastore | Hive Metastore server | <code>\$ sudo chkconfig hive-metastore on</code>     |
| HiveServer2    | HiveServer2           | <code>\$ sudo chkconfig hive-server2 on</code>       |
| Zookeeper      | Zookeeper server      | <code>\$ sudo chkconfig zookeeper-server on</code>   |
| HttpFS         | HttpFS server         | <code>\$ sudo chkconfig hadoop-httpfs on</code>      |

The `update-rc.d` commands to use on Ubuntu and Debian systems are:

| Component      | Server                | Command                                                      |
|----------------|-----------------------|--------------------------------------------------------------|
| Hue            | Hue server            | <code>\$ sudo update-rc.d hue defaults</code>                |
| Oozie          | Oozie server          | <code>\$ sudo update-rc.d oozie defaults</code>              |
| HBase          | HBase master          | <code>\$ sudo update-rc.d hbase-master defaults</code>       |
|                | HBase slave           | <code>\$ sudo update-rc.d hbase-regionserver defaults</code> |
| Hive Metastore | Hive Metastore server | <code>\$ sudo update-rc.d hive-metastore defaults</code>     |
| HiveServer2    | HiveServer2           | <code>\$ sudo update-rc.d hive-server2 defaults</code>       |

| Component | Server           | Command                                                      |
|-----------|------------------|--------------------------------------------------------------|
| Zookeeper | Zookeeper server | <pre>\$ sudo update-rc.d<br/>zookeeper-server defaults</pre> |
| HttpFS    | HttpFS server    | <pre>\$ sudo update-rc.d<br/>hadoop-httpfs defaults</pre>    |

## Stopping Services

- Important:**

- If you use Cloudera Manager, do not use these command-line instructions.
- This information applies specifically to CDH 5.3.x . If you use an earlier version of CDH, see the documentation for that version located at [Cloudera Documentation](#).

Run the following command on every host in the cluster to shut down all Hadoop Common system services that are started by `init` in the cluster:

```
$ for x in `cd /etc/init.d ; ls hadoop-*` ; do sudo service $x stop ; done
```

To verify that no Hadoop processes are running, issue the following command on each host::

```
# ps -aef | grep java
```

You could also use `ps -fu hdfs` and `ps -fu mapred` to confirm that no processes are running as the `hdfs` or `mapred` user, but additional user names are created by the other components in the ecosystem; checking for the "java" string in the `ps` output is an easy way to identify any processes that may still be running.

To stop system services individually, use the instructions in the table below.

STOP system services in this order:

| Order | Service   | Comments | Instructions                                                                                                                                               |
|-------|-----------|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1     | Hue       |          | Run the following on the Hue Server machine to stop Hue<br><code>sudo service hue stop</code>                                                              |
| 2     | Sqoop 1   |          | Run the following on all nodes where it is running:<br><code>sudo service sqoop-metastore stop</code>                                                      |
| 2     | Sqoop 2   |          | Run the following on all nodes where it is running:<br><code>\$ sudo /sbin/service sqoop2-server stop</code>                                               |
| 3     | Flume 0.9 |          | Stop the Flume Node processes on each node where they are running:<br><code>sudo service flume-node stop</code><br>Stop the Flume Master <code>sudo</code> |

| Order | Service      | Comments                                                        | Instructions                                                                                                                                                                                                                                                                                                                                              |
|-------|--------------|-----------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|       |              |                                                                 | <code>service flume-master stop</code>                                                                                                                                                                                                                                                                                                                    |
| 4     | Flume 1.x    | There is no Flume master                                        | <p>Stop the Flume Node processes on each node where they are running:</p> <pre>sudo service flume-ng-agent stop</pre>                                                                                                                                                                                                                                     |
| 5     | Oozie        |                                                                 | <code>sudo service oozie stop</code>                                                                                                                                                                                                                                                                                                                      |
| 6     | Hive         |                                                                 | <p>To stop Hive, exit the Hive console and make sure no Hive scripts are running.</p> <p>Shut down HiveServer2:</p> <pre>sudo service hiveserver2 stop</pre> <p>Shut down the Hive metastore daemon on each client:</p> <pre>sudo service hive-metastore stop</pre> <p>If the metastore is running from the command line, use Ctrl-c to shut it down.</p> |
| 7     | HBase        | Stop the Thrift server and clients, then shut down the cluster. | <p>To stop the Thrift server and clients:</p> <pre>sudo service hbase-thrift stop</pre> <p>To shut down the cluster, use this command on the master node:</p> <pre>sudo service hbase-master stop</pre> <p>Use the following command on each node hosting a region server:</p> <pre>sudo service hadoop-hbase-regionserver stop</pre>                     |
| 8a    | MapReduce v1 | Stop Hive and Oozie before stopping MapReduce.                  | <p>To stop MapReduce, stop the JobTracker service, and stop the Task Tracker on all nodes where it is running. Use the following commands:</p> <pre>sudo service hadoop-0.20-mapreduce-jobtracker stop</pre>                                                                                                                                              |

| Order | Service   | Comments                                       | Instructions                                                                                                                                                                                                                                                                                                                           |
|-------|-----------|------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|       |           |                                                | <pre>sudo service <del>hadoop-0.20-mapreduce-tasktracker</del> stop</pre>                                                                                                                                                                                                                                                              |
| 8b    | YARN      | Stop Hive and Oozie before stopping YARN.      | <p>To stop YARN, stop the MapReduce JobHistory service, ResourceManager service, and NodeManager on all nodes where they are running. Use the following commands:</p> <pre>sudo service <del>hadoop-mapreduce-historyserver</del> stop  sudo service hadoop-yarn-resourcemanager stop  sudo service hadoop-yarn-nodemanager stop</pre> |
| 9     | HttpFS    |                                                | <pre>sudo service hadoop-httpfs stop</pre>                                                                                                                                                                                                                                                                                             |
| 10    | HDFS      |                                                | <p>To stop HDFS: On the NameNode: <code>sudo service hadoop-hdfs-namenode stop</code></p> <p>On the Secondary NameNode (if used): <code>sudo service hadoop-hdfs-secondarynamenode stop</code></p> <p>On each DataNode: <code>sudo service hadoop-hdfs-datanode stop</code></p>                                                        |
| 11    | ZooKeeper | Stop HBase and HDFS before stopping ZooKeeper. | <p>To stop the ZooKeeper server, use one of the following commands on each ZooKeeper node:</p> <pre>sudo service zookeeper-server stop  or  sudo service zookeeper stop</pre>                                                                                                                                                          |

### Migrating Data between a CDH 4 and CDH 5 Cluster

You can migrate the data from a CDH 4 (or any Apache Hadoop) cluster to a CDH 5 cluster by using a tool that copies out data in parallel, such as the DistCp tool offered in CDH 5. This can be useful if you are not planning to upgrade your CDH 4 cluster itself at this point. The following sections provide information and instructions:

- [Requirements and Restrictions](#) on page 118
- [Copying Data between two Clusters Using distcp](#) on page 118
- [Post-Migration Verification](#)

#### Requirements and Restrictions

1. The CDH 5 cluster must have a MapReduce service running on it (MRv1 or YARN (MRv2)).
2. All the MapReduce nodes in the CDH 5 cluster should have full network access to all the nodes of the source cluster. This allows you to perform the copy in a distributed manner.
3. To copy data between a secure and an insecure cluster, you must run the `distcp` command on the secure cluster.
4. To copy data from a CDH 4 to a CDH 5 cluster, you can do one of the following:

- **Note:**

The term **source** in this case refers to the CDH 4 (or other Hadoop) cluster you want to migrate or copy data from; and **destination** refers to the CDH 5 cluster.

- Running commands on the destination cluster, use the Hftp protocol for the source cluster, and HDFS for the destination. (Hftp is read-only, so you must run DistCp on the destination cluster and pull the data from the source cluster.) See [Copying Data between two Clusters Using distcp](#) on page 118.

- **Note:**

Do not use this method if one of the clusters is secure and the other is not.

- Running commands on the source cluster, use the HDFS or webHDFS protocol for the source cluster, and webHDFS for the destination. See [Copying Data between a Secure and an Insecure Cluster using DistCp and webHDFS](#) on page 121.
- Running commands on the destination cluster, use webHDFS for the source cluster, and webHDFS for the destination. See [Copying Data between a Secure and an Insecure Cluster using DistCp and webHDFS](#) on page 121.

The following restrictions currently apply (see [Apache Hadoop Known Issues](#)):

- DistCp does not work between a secure cluster and an insecure cluster in some cases.

As of CDH 5.1.3, DistCp *does* work between a secure and an insecure cluster if you use the webHDFS protocol and run the command from the secure cluster side after setting `ipc.client.fallback-to-simple-auth-allowed` to true, as described under [Copying Data between a Secure and an Insecure Cluster using DistCp and webHDFS](#) on page 121.

- To use DistCp using Hftp from a secure cluster using SPNEGO, you must configure the `dfs.https.port` property on the client to use the HTTP port (50070 by default).

#### Copying Data between two Clusters Using `distcp`

- **Important:** Do not run `distcp` as the HDFS user. The HDFS user is blacklisted for MapReduce jobs by default.

You can use the `distcp` tool on the destination cluster to initiate the copy job to move the data. Between two clusters running different versions of CDH, run the `distcp` tool with `hftp://` as the source file system and

`hdfs://` as the destination file system. This uses the HFTP protocol for the source and the HDFS protocol for the destination. The default port for HFTP is 50070 and the default port for HDFS is 8020. Amazon S3 block and native filesystems are also supported, via `s3://` or `s3n://` protocols.

**Example of a source URI:** `hftp://namenode-location:50070/basePath`

where `namenode-location` refers to the CDH 4 NameNode hostname as defined by its configured `fs.default.name` and 50070 is the NameNode's HTTP server port, as defined by the configured `dfs.http.address`.

**Example of a destination URI:** `hdfs://nameservice-id/basePath` OR `hdfs://namenode-location`

This refers to the CDH 5 NameNode as defined by its configured `fs.defaultFS`.

The `basePath` in both the above URIs refers to the directory you want to copy, if one is specifically needed.

**Example of an Amazon S3 Block Filesystem URI:** `s3://accessKeyId:secretkey@bucket/file`

**Example of an Amazon S3 Native Filesystem URI:** `s3n://accessKeyId:secretkey@bucket/file`

### The `distcp` Command

For more help, and to see all the options available on the `distcp` command, use the following command to see the built-in help:

```
$ hadoop distcp
```

Run the `distcp` copy by issuing a command such as the following on the CDH 5 cluster:

- **Important:** If you are using `distcp` as part of an upgrade, run the following commands on the destination cluster only, in this example, the CDH 5 cluster, during the upgrade.

```
$ hadoop distcp hftp://cdh4-namenode:50070/ hdfs://CDH5-nameservice/
$ hadoop distcp s3://accessKeyId:secretkey@bucket/ hdfs://CDH5-nameservice/
$ hadoop distcp s3n://accessKeyId:secretkey@bucket/ hdfs://CDH5-nameservice/
```

Or use a specific path, such as `/hbase` to move HBase data, for example:

```
$ hadoop distcp hftp://cdh4-namenode:50070/hbase hdfs://CDH5-nameservice/hbase
$ hadoop distcp s3://accessKeyId:secretkey@bucket/file
hdfs://CDH5-nameservice/bucket/file
$ hadoop distcp s3n://accessKeyId:secretkey@bucket/file
hdfs://CDH5-nameservice/bucket/file
```

`distcp` will then submit a regular MapReduce job that performs a file-by-file copy.

`distcp` is a general utility for copying files between distributed filesystems in different clusters. In general, you can use `distcp` to copy files between compatible clusters in either direction. For instance, you can copy files from a HDFS filesystem to an Amazon S3 filesystem. The examples given above are specific to using `distcp` to assist with an upgrade from CDH 4 to CDH 5.

### Protocol Support for `distcp`

The following table lists support for using different protocols with the `distcp` command on different versions of CDH. In the table, **secure** means that the cluster is configured to use Kerberos. Copying between a secure cluster and an insecure cluster is only supported from CDH 5.1.3 onward, due to the inclusion of [HDFS-6776](#).

- **Note:** HFTP is a read-only protocol and cannot be used for the destination.

| Source | Destination              | Source protocol and configuration | Destination protocol and configuration | Where to issue distcp command | Fallback Configuration Required | Status |
|--------|--------------------------|-----------------------------------|----------------------------------------|-------------------------------|---------------------------------|--------|
| CDH 4  | CDH 4                    | hdfs or webhdfs, insecure         | hdfs or webhdfs, insecure              | Source or destination         |                                 | ok     |
| CDH 4  | CDH 4                    | hftp, insecure                    | hdfs or webhdfs, insecure              | Destination                   |                                 | ok     |
| CDH 4  | CDH 4                    | hdfs or webhdfs, secure           | hdfs or webhdfs, secure                | Source or destination         |                                 | ok     |
| CDH 4  | CDH 4                    | hftp, secure                      | hdfs or webhdfs, secure                | Destination                   |                                 | ok     |
| CDH 4  | CDH 5                    | webhdfs, insecure                 | webhdfs or hdfs, insecure              | Destination                   |                                 | ok     |
| CDH 4  | CDH 5 ( 5.1.3 and newer) | webhdfs, insecure                 | webhdfs, secure                        | Destination                   | yes                             | ok     |
| CDH 4  | CDH 5                    | webhdfs or hftp, insecure         | webhdfs or hdfs, insecure              | Destination                   |                                 | ok     |
| CDH 4  | CDH 5                    | webhdfs or hftp, secure           | webhdfs or hdfs, secure                | Destination                   |                                 | ok     |
| CDH 5  | CDH 4                    | webhdfs , insecure                | webhdfs, insecure                      | source or destination         |                                 | ok     |
| CDH 5  | CDH 4                    | webhdfs , insecure                | hdfs, insecure                         | destination                   |                                 | ok     |
| CDH 5  | CDH 4                    | hdfs, insecure                    | webhdfs, insecure                      | source                        |                                 | ok     |
| CDH 5  | CDH 4                    | hftp, insecure                    | hdfs or webhdfs, insecure              | destination                   |                                 | ok     |
| CDH 5  | CDH 4                    | webhdfs, secure                   | webhdfs, secure                        | source or destination         |                                 | ok     |
| CDH 5  | CDH 4                    | webhdfs, secure                   | hdfs, insecure                         | destination                   |                                 | ok     |
| CDH 5  | CDH 4                    | hdfs, secure                      | webhdfs, secure                        | source                        |                                 | ok     |
| CDH 5  | CDH 5                    | hdfs or webhdfs, insecure         | hdfs or webhdfs, insecure              | Source or destination         |                                 | ok     |
| CDH 5  | CDH 5                    | hftp, insecure                    | hdfs or webhdfs, insecure              | Destination                   |                                 | ok     |



| Source | Destination             | Source protocol and configuration | Destination protocol and configuration | Where to issue distcp command | Fallback Configuration Required | Status |
|--------|-------------------------|-----------------------------------|----------------------------------------|-------------------------------|---------------------------------|--------|
| CDH 5  | CDH 5                   | hdfs or webhdfs, secure           | hdfs or webhdfs, secure                | Source or destination         |                                 | ok     |
| CDH 5  | CDH 5                   | hftp, secure                      | hdfs or webhdfs, secure                | Destination                   |                                 | ok     |
| CDH 5  | CDH 5 (5.1.3 and newer) | hdfs or webhdfs, secure           | hdfs or webhdfs, insecure              | Source                        | yes                             | ok     |

To enable the fallback configuration, for copying between a secure cluster and an insecure one, add the following to the HDFS `core-default.xml`, by using an advanced configuration snippet if you use Cloudera Manager, or editing the file directly otherwise.

```
<property>
  <name>ipc.client.fallback-to-simple-auth-allowed</name>
  <value>true</value>
</property>
```

## Copying Data between a Secure and an Insecure Cluster using DistCp and webHDFS

You can use DistCp and webHDFS to copy data between a secure cluster and an insecure cluster by doing the following:

1. Set `ipc.client.fallback-to-simple-auth-allowed` to true in `core-site.xml` *on the secure cluster side*:

```
<property>
  <name>ipc.client.fallback-to-simple-auth-allowed</name>
  <value>true</value>
</property>
```

2. Use commands such as the following *from the secure cluster side only*:

```
distcp webhdfs://insecureCluster webhdfs://secureCluster
distcp webhdfs://secureCluster webhdfs://insecureCluster
```

## Post-migration Verification

After migrating data between the two clusters, it is a good idea to use `hadoop fs -ls /basePath` to verify the permissions, ownership and other aspects of your files, and correct any problems before using the files in your new cluster.

## Configuring HDFS

### Configuring an NFSv3 Gateway

- **Important:**

- If you use Cloudera Manager, do not use these command-line instructions.
- This information applies specifically to CDH 5.3.x. If you use an earlier version of CDH, see the documentation for that version located at [Cloudera Documentation](#).

## Configuring CDH and Managed Services

The NFSv3 gateway allows a client to mount HDFS as part of the client's local file system. The gateway machine can be any host in the cluster, including the NameNode, a DataNode, or any HDFS client. The client can be any NFSv3-client-compatible machine.

After mounting HDFS to his or her local filesystem, a user can:

- Browse the HDFS file system as though it were part of the local file system
- Upload and download files from the HDFS file system to and from the local file system.
- Stream data directly to HDFS through the mount point.

File append is supported, but random write is not.

The subsections that follow provide information on installing and configuring the gateway.

### ■ **Note: Install Cloudera Repository**

Before using the instructions on this page to install or upgrade, install the Cloudera `yum`, `zypper`/`YaST` or `apt` repository, and install or upgrade CDH 5 and make sure it is functioning correctly. For instructions, see [Installing the Latest CDH 5 Release](#) and [Upgrading Unmanaged CDH Using the Command Line](#).

### Upgrading from a CDH 5 Beta Release

If you are upgrading from a CDH 5 Beta release, you must first remove the `hadoop-hdfs-portmap` package. Proceed as follows.

1. Unmount existing HDFS gateway mounts. For example, on each client, assuming the file system is mounted on `/hdfs_nfs_mount`:

```
$ umount /hdfs_nfs_mount
```

2. Stop the services:

```
$ sudo service hadoop-hdfs-nfs3 stop
$ sudo hadoop-hdfs-portmap stop
```

3. Remove the `hadoop-hdfs-portmap` package.

- On a RHEL-compatible system:

```
$ sudo yum remove hadoop-hdfs-portmap
```

- On a SLES system:

```
$ sudo zypper remove hadoop-hdfs-portmap
```

- On an Ubuntu or Debian system:

```
$ sudo apt-get remove hadoop-hdfs-portmap
```

4. Install the new version

- On a RHEL-compatible system:

```
$ sudo yum install hadoop-hdfs-nfs3
```

- On a SLES system:

```
$ sudo zypper install hadoop-hdfs-nfs3
```

- On an Ubuntu or Debian system:

```
$ sudo apt-get install hadoop-hdfs-nfs3
```

5. Start the system default portmapper service:

```
$ sudo service portmap start
```

6. Now proceed with [Starting the NFSv3 Gateway](#) on page 124, and then [remount the HDFS gateway mounts](#).

### Installing the Packages for the First Time

#### On RHEL and similar systems:

Install the following packages on the cluster host you choose for NFSv3 Gateway machine (we'll refer to it as the NFS server from here on).

- nfs-utils
- nfs-utils-lib
- hadoop-hdfs-nfs3

The first two items are standard NFS utilities; the last is a CDH package.

Use the following command:

```
$ sudo yum install nfs-utils nfs-utils-lib hadoop-hdfs-nfs3
```

#### On SLES:

Install `nfs-utils` on the cluster host you choose for NFSv3 Gateway machine (referred to as the NFS server from here on):

```
$ sudo zypper install nfs-utils
```

#### On an Ubuntu or Debian system:

Install `nfs-common` on the cluster host you choose for NFSv3 Gateway machine (referred to as the NFS server from here on):

```
$ sudo apt-get install nfs-common
```

### Configuring the NFSv3 Gateway

Proceed as follows to configure the gateway.

1. Add the following property to `hdfs-site.xml` on the *NameNode*:

```
<property>
  <name>dfs.namenode.accesstime.precision</name>
  <value>3600000</value>
  <description>The access time for an HDFS file is precise up to this value. The
    default value is 1 hour.
    Setting a value of 0 disables access times for HDFS.</description>
</property>
```

2. Add the following property to `hdfs-site.xml` on the *NFS server*:

```
<property>
  <name>dfs.nfs3.dump.dir</name>
  <value>/tmp/.hdfs-nfs</value>
</property>
```

■ **Note:**

You should change the location of the file dump directory, which temporarily saves out-of-order writes before writing them to HDFS. This directory is needed because the NFS client often reorders writes, and so sequential writes can arrive at the NFS gateway in random order and need to be saved until they can be ordered correctly. After these out-of-order writes have exceeded 1MB in memory for any given file, they are dumped to the `dfs.nfs3.dump.dir` (the memory threshold is not currently configurable).

Make sure the directory you choose has enough space. For example, if an application uploads 10 files of 100MB each, `dfs.nfs3.dump.dir` should have roughly 1GB of free space to allow for a worst-case reordering of writes to every file.

3. Configure the user running the gateway (normally the `hdfs` user as in this example) to be a proxy for other users. To allow the `hdfs` user to be a proxy for all other users, add the following entries to `core-site.xml` on the NameNode:

```
<property>
  <name>hadoop.proxyuser.hdfs.groups</name>
  <value>*</value>
  <description>
    Set this to '*' to allow the gateway user to proxy any group.
  </description>
</property>
<property>
  <name>hadoop.proxyuser.hdfs.hosts</name>
  <value>*</value>
  <description>
    Set this to '*' to allow requests from any hosts to be proxied.
  </description>
</property>
```

4. Restart the NameNode.

### Starting the NFSv3 Gateway

Do the following on the NFS server.

1. First, stop the default NFS services, if they are running:

```
$ sudo service nfs stop
```

2. Start the HDFS-specific services:

```
$ sudo service hadoop-hdfs-nfs3 start
```

### Verifying that the NFSv3 Gateway is Working

To verify that the NFS services are running properly, you can use the `rpcinfo` command on any host on the local network:

```
$ rpcinfo -p <nfs_server_ip_address>
```

You should see output such as the following:

program	vers	proto	port	
100005	1	tcp	4242	mountd
100005	2	udp	4242	mountd
100005	2	tcp	4242	mountd
100000	2	tcp	111	portmapper
100000	2	udp	111	portmapper

```
100005      3      udp      4242  mountd
100005      1      udp      4242  mountd
100003      3      tcp      2049  nfs
100005      3      tcp      4242  mountd
```

To verify that the HDFS namespace is exported and can be mounted, use the `showmount` command.

```
$ showmount -e <nfs_server_ip_address>
```

You should see output similar to the following:

```
Exports list on <nfs_server_ip_address>:
/ (everyone)
```

### Mounting HDFS on an NFS Client

To import the HDFS file system on an NFS client, use a `mount` command such as the following on the client:

```
$ mount -t nfs -o vers=3,proto=tcp,nolock <nfs_server_hostname>:/ /hdfs_nfs_mount
```

#### Note:

When you create a file or directory as user `hdfs` on the client (that is, in the HDFS file system imported via the NFS mount), the ownership may differ from what it would be if you had created it in HDFS directly. For example, ownership of a file created on the client might be `hdfs:hdfs` when the same operation done natively in HDFS resulted in `hdfs:supergroup`. This is because in native HDFS, BSD semantics determine the group ownership of a newly-created file: it is set to the same group as the parent directory where the file is created. When the operation is done over NFS, the typical Linux semantics create the file with the group of the effective GID (group ID) of the process creating the file, and this characteristic is explicitly passed to the NFS gateway and HDFS.

### Configuring Mountable HDFS

CDH 5 includes a FUSE (Filesystem in Userspace) interface into HDFS. The `hadoop-hdfs-fuse` package enables you to use your HDFS cluster as if it were a traditional filesystem on Linux. Proceed as follows.

**Before you start:** You must have a working HDFS cluster and know the hostname and port that your NameNode exposes.

**To install `hadoop-hdfs-fuses` On Red Hat-compatible systems:**

```
$ sudo yum install hadoop-hdfs-fuse
```

**To install `hadoop-hdfs-fuse` on Ubuntu systems:**

```
$ sudo apt-get install hadoop-hdfs-fuse
```

**To install `hadoop-hdfs-fuse` on SLES systems:**

```
$ sudo zypper install hadoop-hdfs-fuse
```

You now have everything you need to begin mounting HDFS on Linux.

**To set up and test your mount point in a non-HA installation:**

```
$ mkdir -p <mount_point>
$ hadoop-fuse-dfs dfs://<name_node_hostname>:<namenode_port> <mount_point>
```

where `namenode_port` is the NameNode's RPC port, `dfs.namenode.servicerpc-address`.

### To set up and test your mount point in an HA installation:

```
$ mkdir -p <mount_point>
$ hadoop-fuse-dfs dfs://<nameservice_id> <mount_point>
```

where *nameservice\_id* is the value of *fs.defaultFS*. In this case the port defined for *dfs.namenode.rpc-address.[nameservice ID].[name node ID]* is used automatically. See [Enabling HDFS HA](#) on page 213 for more information about these properties.

You can now run operations as if they are on your mount point. Press Ctrl+C to end the *fuse-dfs* program, and *umount* the partition if it is still mounted.

#### ■ Note:

To find its configuration directory, *hadoop-fuse-dfs* uses the *HADOOP\_CONF\_DIR* configured at the time the *mount* command is invoked.

### To clean up your test:

```
$ umount <mount_point>
```

You can now add a permanent HDFS mount which persists through reboots.

### To add a system mount:

1. Open */etc/fstab* and add lines to the bottom similar to these:

```
hadoop-fuse-dfs#dfs://<name_node_hostname>:<namenode_port> <mount_point> fuse
allow_other,usetrash,rw 2 0
```

For example:

```
hadoop-fuse-dfs#dfs://localhost:8020 /mnt/hdfs fuse allow_other,usetrash,rw 2 0
```

#### ■ Note:

In an HA deployment, use the HDFS nameservice instead of the NameNode URI; that is, use the value of *dfs.nameservices* in *hdfs-site.xml*.

2. Test to make sure everything is working properly:

```
$ mount <mount_point>
```

Your system is now configured to allow you to use the *ls* command and use that mount point as if it were a normal system disk.

By default, the CDH 5 package installation creates the */etc/default/hadoop-fuse* file with a maximum heap size of 128 MB. You can change the JVM minimum and maximum heap size; for example

To change it:

```
export LIBHDFS_OPTS="-Xms64m -Xmx256m"
```

Be careful not to set the minimum to a higher value than the maximum.

For more information, see the help for *hadoop-fuse-dfs*:

```
$ hadoop-fuse-dfs --help
```

## Using the HDFS Balancer

### Important:

- If you use Cloudera Manager, do not use these command-line instructions.
- This information applies specifically to CDH 5.3.x. If you use an earlier version of CDH, see the documentation for that version located at [Cloudera Documentation](#).

The HDFS balancer re-balances data across the DataNodes, moving blocks from over-utilized to under-utilized nodes. As the system administrator, you can run the balancer from the command-line as necessary -- for example, after adding new DataNodes to the cluster.

Points to note:

- The balancer requires the capabilities of an HDFS superuser (for example, the `hdfs` user) to run.
- The balancer does not balance between individual volumes on a single DataNode.
- You can run the balancer without parameters, as follows:

```
sudo -u hdfs hdfs balancer
```

### Note:

If [Kerberos is enabled](#), do not use commands in the form `sudo -u <user> hadoop <command>`; they will fail with a security error. Instead, use the following commands: `$ kinit <user>` (if you are using a password) *or* `$ kinit -kt <keytab> <principal>` (if you are using a keytab) and then, for each command executed by this user, `$ <command>`

This runs the balancer with a default threshold of 10%, meaning that the script will ensure that disk usage on each DataNode differs from the overall usage in the cluster by no more than 10%. For example, if overall usage across all the DataNodes in the cluster is 40% of the cluster's total disk-storage capacity, the script ensures that each DataNode's disk usage is between 30% and 50% of that DataNode's disk-storage capacity.

- You can run the script with a different threshold; for example:

```
sudo -u hdfs hdfs balancer -threshold 5
```

This specifies that each DataNode's disk usage must be (or will be adjusted to be) within 5% of the cluster's overall usage.

- You can adjust the network bandwidth used by the balancer, by running the `dfsadmin -setBalancerBandwidth` command before you run the balancer; for example:

```
dfsadmin -setBalancerBandwidth newbandwidth
```

where *newbandwidth* is the maximum amount of network bandwidth, in bytes per second, that each DataNode can use during the balancing operation. For more information about the bandwidth command, see [this page](#).

- The balancer can take a long time to run, especially if you are running it for the first time, or do not run it regularly.

## Setting HDFS Quotas

You can set quotas in HDFS for:

- The number of file and directory names used
- The amount of space used by given directories

Points to note:

- The quotas for names and the quotas for space are independent of each other.
- File and directory creation fails if the creation would cause the quota to be exceeded.

## Configuring CDH and Managed Services

- Allocation fails if the quota would prevent a full block from being written; keep this in mind if you are using a large block size.
- If you are using replication, remember that each replica of a block counts against the quota.

### Commands

- **Important:**

- If you use Cloudera Manager, do not use these command-line instructions.
- This information applies specifically to CDH 5.3.x . If you use an earlier version of CDH, see the documentation for that version located at [Cloudera Documentation](#).

To set space quotas on a directory:

```
dfsadmin -setSpaceQuota n directory
```

where *n* is a number of bytes and *directory* is the directory the quota applies to. You can specify multiple directories in a single command; *n* applies to each.

To remove space quotas from a directory:

```
dfsadmin -clrSpaceQuota directory
```

You can specify multiple directories in a single command.

To set name quotas on a directory:

```
dfsadmin -setQuota n directory
```

where *n* is the number of file and directory names in *directory*. You can specify multiple directories in a single command; *n* applies to each.

To remove name quotas from a directory:

```
dfsadmin -clrQuota directory
```

You can specify multiple directories in a single command.

### For More Information

For more information, see the [HDFS Quotas Guide](#).

## Configuring Centralized Cache Management in HDFS

Centralized cache management in HDFS is an explicit caching mechanism that allows users to specify paths to be cached by HDFS. The NameNode will communicate with DataNodes that have the desired blocks on disk, and instruct them to cache the blocks in off-heap caches.

This has several advantages:

- Explicit pinning prevents frequently used data from being evicted from memory. This is particularly important when the size of the working set exceeds the size of main memory, which is common for many HDFS workloads.
- Since DataNode caches are managed by the NameNode, applications can query the set of cached block locations when making task placement decisions. Co-locating a task with a cached block replica improves read performance.
- When block has been cached by a DataNode, clients can use a new, more-efficient, zero-copy read API. Since checksum verification of cached data is done once by the DataNode, clients can incur essentially zero overhead when using this new API.



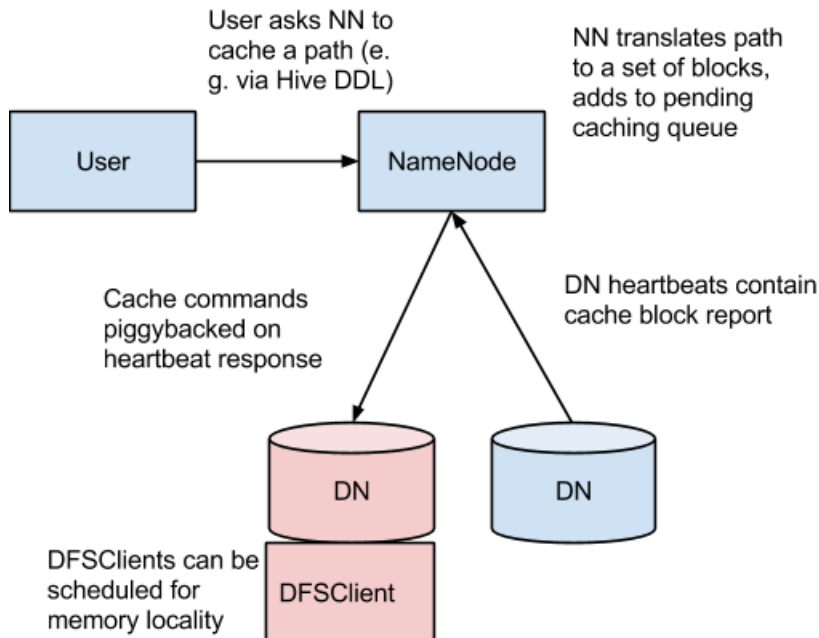
- Centralized caching can improve overall cluster memory utilization. When relying on the OS buffer cache at each DataNode, repeated reads of a block will result in all  $n$  replicas of the block being pulled into buffer cache. With centralized cache management, you can explicitly pin only  $m$  of the  $n$  replicas, saving  $n-m$  memory.

### Use Cases

Centralized cache management is most useful for files that are accessed repeatedly. For example, a fact table in Hive which is often used in joins is a good candidate for caching. On the other hand, caching the input of a one-year reporting query is probably less useful, since the historical data might be read only once.

Centralized cache management is also useful for mixed workloads with performance SLAs. Caching the working set of a high-priority workload insures that it does not contend for disk I/O with a low-priority workload.

### Architecture



In this architecture, the NameNode is responsible for coordinating all the DataNode off-heap caches in the cluster. The NameNode periodically receives a "cache report" from each DataNode which describes all the blocks cached on a given DataNode. The NameNode manages DataNode caches by piggybacking cache and uncache commands on the DataNode heartbeat.

The NameNode queries its set of cache directives to determine which paths should be cached. Cache directives are persistently stored in the fsimage and edit log, and can be added, removed, and modified using Java and command-line APIs. The NameNode also stores a set of cache pools, which are administrative entities used to group cache directives together for resource management and enforcing permissions.

The NameNode periodically rescans the namespace and active cache directories to determine which blocks need to be cached or uncached and assign caching to DataNodes. Rescans can also be triggered by user actions like adding or removing a cache directive or removing a cache pool.

We do not currently cache blocks which are under construction, corrupt, or otherwise incomplete. If a cache directive covers a symlink, the symlink target is not cached. Caching is currently done on a per-file basis, although we would like to add block-level granularity in the future.

### Concepts

#### Cache Directive

A **cache directive** defines a path that should be cached. Paths can be either directories or files. Directories are cached non-recursively, meaning only files in the first-level listing of the directory will be cached. Directives also specify additional parameters, such as the cache replication factor and expiration time. The replication factor

specifies the number of block replicas to cache. If multiple cache directives refer to the same file, the maximum cache replication factor is applied.

The expiration time is specified on the command line as a `time-to-live` (TTL), a relative expiration time in the future. After a cache directive expires, it is no longer considered by the NameNode when making caching decisions.

### Cache Pool

A **cache pool** is an administrative entity used to manage groups of cache directives. Cache pools have UNIX-like permissions that restrict which users and groups have access to the pool. Write permissions allow users to add and remove cache directives to the pool. Read permissions allow users to list the cache directives in a pool, as well as additional metadata. Execute permissions are unused.

Cache pools are also used for resource management. Pools can enforce a maximum `limit`, which restricts the number of bytes that can be cached in aggregate by directives in the pool. Normally, the sum of the pool limits will approximately equal the amount of aggregate memory reserved for HDFS caching on the cluster. Cache pools also track a number of statistics to help cluster users determine what is and should be cached.

Pools also enforce a maximum time-to-live. This restricts the maximum expiration time of directives being added to the pool.

### cacheadmin Command-Line Interface

#### ■ Important:

- If you use Cloudera Manager, do not use these command-line instructions.
- This information applies specifically to CDH 5.3.x. If you use an earlier version of CDH, see the documentation for that version located at [Cloudera Documentation](#).

On the command-line, administrators and users can interact with cache pools and directives via the `hdfs cacheadmin` subcommand. Cache directives are identified by a unique, non-repeating 64-bit integer ID. IDs will not be reused even if a cache directive is later removed. Cache pools are identified by a unique string name.

### Cache Directive Commands

#### addDirective

**Description:** Add a new cache directive.

**Usage:** `hdfs cacheadmin -addDirective -path <path> -pool <pool-name> [-force] [-replication <replication>] [-ttl <time-to-live>]`

Where, `path`: A path to cache. The path can be a directory or a file.

`pool-name`: The pool to which the directive will be added. You must have write permission on the cache pool in order to add new directives.

`force`: Skips checking of cache pool resource limits.

`replication`: The cache replication factor to use. Defaults to 1.

`time-to-live`: Time period for which the directive is valid. Can be specified in seconds, minutes, hours, and days, for example: 30m, 4h, 2d. The value `never` indicates a directive that never expires. If unspecified, the directive never expires.

#### removeDirective

**Description:** Remove a cache directive.

**Usage:** `hdfs cacheadmin -removeDirective <id>`

Where, `id`: The id of the cache directive to remove. You must have write permission on the pool of the directive in order to remove it. To see a list of PathBasedCache directive IDs, use the `-listDirectives` command.

### removeDirectives

**Description:** Remove every cache directive with the specified path.

**Usage:** `hdfs cacheadmin -removeDirectives <path>`

Where, `path`: The path of the cache directives to remove. You must have write permission on the pool of the directive in order to remove it.

### listDirectives

**Description:** List PathBasedCache directives.

**Usage:** `hdfs cacheadmin -listDirectives [-stats] [-path <path>] [-pool <pool>]`

Where, `path`: List only PathBasedCache directives with this path. Note that if there is a PathBasedCache directive for `path` in a cache pool that we don't have read access for, it will not be listed.

`pool`: List only path cache directives in that pool.

`stats`: List path-based cache directive statistics.

### Cache Pool Commands

#### addPool

**Description:** Add a new cache pool.

**Usage:** `hdfs cacheadmin -addPool <name> [-owner <owner>] [-group <group>] [-mode <mode>] [-limit <limit>] [-maxTtl <maxTtl>]`

Where, `name`: Name of the new pool.

`owner`: Username of the owner of the pool. Defaults to the current user.

`group`: Group of the pool. Defaults to the primary group name of the current user.

`mode`: UNIX-style permissions for the pool. Permissions are specified in octal, for example: 0755. By default, this is set to 0755.

`limit`: The maximum number of bytes that can be cached by directives in this pool, in aggregate. By default, no limit is set.

`maxTtl`: The maximum allowed time-to-live for directives being added to the pool. This can be specified in seconds, minutes, hours, and days, for example: 120s, 30m, 4h, 2d. By default, no maximum is set. A value of `never` specifies that there is no limit.

#### modifyPool

**Description:** Modifies the metadata of an existing cache pool.

**Usage:** `hdfs cacheadmin -modifyPool <name> [-owner <owner>] [-group <group>] [-mode <mode>] [-limit <limit>] [-maxTtl <maxTtl>]`

Where, `name`: Name of the pool to modify.

`owner`: Username of the owner of the pool.

`group`: Groupname of the group of the pool.

`mode`: Unix-style permissions of the pool in octal.

`limit`: Maximum number of bytes that can be cached by this pool.

`maxTtl`: The maximum allowed time-to-live for directives being added to the pool.

### removePool

**Description:** Remove a cache pool. This also uncaches paths associated with the pool.

**Usage:** `hdfs cacheadmin -removePool <name>`

Where, `name`: Name of the cache pool to remove.

### listPools

**Description:** Display information about one or more cache pools, for example: name, owner, group, permissions, and so on.

**Usage:** `hdfs cacheadmin -listPools [-stats] [<name>]`

Where, `name`: If specified, list only the named cache pool.

`stats`: Display additional cache pool statistics.

### help

**Description:** Get detailed help about a command.

**Usage:** `hdfs cacheadmin -help <command-name>`

Where, `command-name`: The command for which to get detailed help. If no command is specified, print detailed help for all commands.

## Configuration

### Native Libraries

In order to lock block files into memory, the DataNode relies on native JNI code found in `libhadoop.so`. Be sure to [enable JNI](#) if you are using HDFS centralized cache management.

### Configuration Properties

#### Required

Be sure to configure the following:

- `dfs.datanode.max.locked.memory`: The maximum amount of memory a DataNode will use for caching (in bytes). The "locked-in-memory size" `ulimit (ulimit -l)` of the DataNode user also needs to be increased to match this parameter (see [OS Limits](#)). When setting this value, remember that you will need space in memory for other things as well, such as the DataNode and application JVM heaps and the operating system page cache.

#### Optional

The following properties are not required, but may be specified for tuning:

- `dfs.namenode.path.based.cache.refresh.interval.ms`: The NameNode uses this as the amount of milliseconds between subsequent path cache rescans. This calculates the blocks to cache and each DataNode containing a replica of the block that should cache it. By default, this parameter is set to 300000, which is five minutes.
- `dfs.datanode.fsdatasetcache.max.threads.per.volume`: The DataNode uses this as the maximum number of threads per volume to use for caching new data. By default, this parameter is set to 4.
- `dfs.cachereport.intervalMsec`: The DataNode uses this as the amount of milliseconds between sending a full report of its cache state to the NameNode. By default, this parameter is set to 10000, which is 10 seconds.
- `dfs.namenode.path.based.cache.block.map.allocation.percent`: The percentage of the Java heap which we will allocate to the cached blocks map. The cached blocks map is a hash map which uses chained hashing. Smaller maps may be accessed more slowly if the number of cached blocks is large; larger maps will consume more memory. By default, this parameter is set to 0.25 percent.

### OS Limits

If you get the error `Cannot start datanode because the configured max locked memory size... is more than the datanode's available RLIMIT_MEMLOCK ulimit`, that means that the operating system is imposing a lower limit on the amount of memory that you can lock than what you have configured. To fix this, you must adjust the `ulimit -l` value that the DataNode runs with. Usually, this value is configured in `/etc/security/limits.conf`. However, it will vary depending on what operating system and distribution you are using.

You will know that you have correctly configured this value when you can run `ulimit -l` from the shell and get back either a higher value than what you have configured with `dfs.datanode.max.locked.memory`, or the string `unlimited`, indicating that there is no limit. Note that it's typical for `ulimit -l` to output the memory lock limit in KB, but `dfs.datanode.max.locked.memory` must be specified in bytes.

## Configuring and Using HBase

### Configuring the Blocksize for HBase

The blocksize is an important configuration option for HBase. HBase data is stored in one (after a major compaction) or more (possibly before a major compaction) HFiles per column family per region. It determines both of the following:

- The blocksize for a given column family determines the smallest unit of data HBase can read from the column family's HFiles.
- It is also the basic unit of measure cached by a RegionServer in the BlockCache.

The default blocksize is 64 KB. The appropriate blocksize is dependent upon your data and usage patterns. Use the following guidelines to tune the blocksize size, in combination with testing and benchmarking as appropriate.

- **Warning:** The default blocksize is appropriate for a wide range of data usage patterns, and tuning the blocksize is an advanced operation. The wrong configuration can negatively impact performance.

- Consider the average key/value size for the column family when tuning the blocksize. You can find the average key/value size using the HFile utility:

```
$ hbase org.apache.hadoop.hbase.io.hfile.HFile -f /path/to/HFILE -m -v
...
Block index size as per heapsize: 296
reader=hdfs://srv1.example.com:9000/path/to/HFILE, \
compression=none, inMemory=false, \
firstKey=US6683275_20040127/mimetype:/1251853756871/Put, \
lastKey=US6684814_20040203/mimetype:/1251864683374/Put, \
avgKeyLen=37, avgValueLen=8, \
entries=1554, length=84447
...
```

- Consider the pattern of reads to the table or column family. For instance, if it is common to scan for 500 rows on various parts of the table, performance might be increased if the blocksize is large enough to encompass 500-1000 rows, so that often, only one read operation on the HFile is required. If your typical scan size is only 3 rows, returning 500-1000 rows would be overkill.

It is difficult to predict the size of a row before it is written, because the data will be compressed when it is written to the HFile. Perform testing to determine the correct blocksize for your data.

### Configuring the Blocksize for a Column Family

You can configure the blocksize of a column family at table creation or by disabling and altering an existing table. These instructions are valid whether or not you use Cloudera Manager to manage your cluster.

```
hbase> create 'test_table',{NAME => 'test_cf', BLOCKSIZE => '262144'}
hbase> disable 'test_table'
```

```
hbase> alter 'test_table', {NAME => 'test_cf', BLOCKSIZE => '524288'}  
hbase> enable 'test_table'
```

After changing the blocksize, the HFiles will be rewritten during the next major compaction. To trigger a major compaction, issue the following command in HBase Shell.

```
hbase> major_compact 'test_table'
```

Depending on the size of the table, the major compaction can take some time and have a performance impact while it is running.

### Monitoring Blocksize Metrics

Several metrics are exposed for monitoring the blocksize by monitoring the blockcache itself. See the `block_cache*` entries in [RegionServer Metrics](#).

### Reading Data from HBase

[Get](#) and [Scan](#) are the two ways to read data from HBase, aside from manually parsing HFiles. A `Get` is simply a `Scan` limited by the API to one row. A `Scan` fetches zero or more rows of a table. By default, a `Scan` reads the entire table from start to end. You can limit your `Scan` results in several different ways, which affect the `Scan`'s load in terms of IO, network, or both, as well as processing load on the client side. This topic is provided as a quick reference. Refer to the [API documentation for Scan](#) for more in-depth information. You can also perform Gets and Scan using the HBase Shell.

- Specify a `startrow` and/or `stoprow`. Neither `startrow` nor `stoprow` need to exist. Because HBase sorts rows lexicographically, it will return the first row after `startrow` would have occurred, and will stop returning rows after `stoprow` would have occurred. The goal is to reduce IO and network.
  - The `startrow` is inclusive and the `stoprow` is exclusive. Given a table with rows `a, b, c, d, e, f`, and `startrow` of `c` and `stoprow` of `f`, rows `c–e` are returned.
  - If you omit `startrow`, the first row of the table is the `startrow`.
  - If you omit the `stoprow`, all results after `startrow` (including `startrow`) are returned.
  - If `startrow` is lexicographically after `stoprow`, and you set `Scan setReversed(boolean reversed)` to `true`, the results are returned in reverse order. Given the same table above, with rows `a–f`, if you specify `c` as the `stoprow` and `f` as the `startrow`, rows `f, e, and d` are returned.

```
Scan()  
Scan(byte[] startRow)  
Scan(byte[] startRow, byte[] stopRow)
```

- Specify a scanner cache that will be filled before the `Scan` result is returned, setting `setCaching` to the number of rows to cache before returning the result. By default, the caching setting on the table is used. The goal is to balance IO and network load.

```
public Scan setCaching(int caching)
```

- To limit the number of columns if your table has very wide rows (rows with a large number of columns), use `setBatch(int batch)` and set it to the number of columns you want to return in one batch. A large number of columns is not a recommended design pattern.

```
public Scan setBatch(int batch)
```

- To specify a maximum result size, use `setMaxResultSize(long)`, with the number of bytes. The goal is to reduce IO and network.

```
public Scan setMaxResultSize(long maxResultSize)
```

- When you use `setCaching` and `setMaxResultSize` together, single server requests are limited by either number of rows or maximum result size, whichever limit comes first.
- You can limit the scan to specific column families or columns by using `addColumn` or `addColumn`. The goal is to reduce IO and network. IO is reduced because each column family is represented by a Store on each region server, and only the Stores representing the specific column families in question need to be accessed.

```
public Scan addColumn(byte[] family,
                     byte[] qualifier)

public Scan addFamily(byte[] family)
```

- You can specify a range of timestamps or a single timestamp by specifying `setTimeRange` or `setTimestamp`.

```
public Scan setTimeRange(long minStamp,
                       long maxStamp)
    throws IOException

public Scan setTimestamp(long timestamp)
    throws IOException
```

- You can retrieve a maximum number of versions by using `setMaxVersions`.

```
public Scan setMaxVersions(int maxVersions)
```

- You can use a filter by using `setFilter`. Filters are discussed in detail in [HBase Filtering](#) on page 136 and the [Filter API](#).

```
public Scan setFilter(Filter filter)
```

- You can disable the server-side block cache for a specific scan by executing `setCacheBlocks(boolean)`. This is an expert setting and should only be used if you know what you are doing.

### Perform Scans Using HBase Shell

You can perform scans using HBase Shell, for testing or quick queries. Use the following guidelines or issue the scan command in HBase Shell with no parameters for more usage information. This represents only a subset of possibilities.

```
# Display usage information
hbase> scan

# Scan all rows of table 't1'
hbase> scan 't1'

# Specify a startrow, limit the result to 10 rows, and only return selected columns
hbase> scan 't1', {COLUMNS => ['c1', 'c2'], LIMIT => 10, STARTROW => 'xyz'}

# Specify a timerange
hbase> scan 't1', {TIMERANGE => [1303668804, 1303668904]}

# Specify a custom filter
hbase> scan 't1', {FILTER => org.apache.hadoop.hbase.filter.ColumnPaginationFilter.new(1,
0)}

# Disable the block cache for a specific scan (experts only)
hbase> scan 't1', {COLUMNS => ['c1', 'c2'], CACHE_BLOCKS => false}
```

### Hedged Reads

Hadoop 2.4 introduced a new feature called *hedged reads*, in [HDFS-5776](#). If a read from a block is slow, the HDFS client starts up another parallel, 'hedged' read against a different block replica. The result of whichever read returns first is used, and the outstanding read is cancelled. This feature helps in situations where a read occasionally takes a long time rather than when there is a systemic problem. Hedged reads can be enabled for HBase when the HFiles are stored in HDFS. This feature is disabled by default.

### Enabling Hedged Reads for HBase Using the Command Line

#### ■ Important:

- If you use Cloudera Manager, do not use these command-line instructions.
- This information applies specifically to CDH 5.3.x. If you use an earlier version of CDH, see the documentation for that version located at [Cloudera Documentation](#).

To enable hedged reads for HBase, edit the `hbase-site.xml` file on each server. Set `dfs.client.hedged.read.threadpool.size` to the number of threads to dedicate to running hedged threads, and set the `dfs.client.hedged.read.threshold.millis` configuration property to the number of milliseconds to wait before starting a second read against a different block replica. Set `dfs.client.hedged.read.threadpool.size` to 0 or remove it from the configuration to disable the feature. After changing these properties, restart your cluster.

The following is an example configuration for hedged reads for HBase.

```
<property>
  <name>dfs.client.hedged.read.threadpool.size</name>
  <value>20</value>  <!-- 20 threads -->
</property>
<property>
  <name>dfs.client.hedged.read.threshold.millis</name>
  <value>10</value>  <!-- 10 milliseconds -->
</property>
```

### HBase Filtering

When reading data from HBase using Get or Scan operations, you can use custom filters to return a subset of results to the client. While this does not reduce server-side IO, it does reduce network bandwidth and reduces the amount of data the client needs to process. Filters are generally used via the Java API, but can be used from HBase Shell for testing and debugging purposes.

For more information on Gets and Scans in HBase, see [Reading Data from HBase](#) on page 134.

#### Filter Syntax Guidelines

HBase filters take zero or more arguments, in parentheses. Where the argument is a string, it is surrounded by single quotes ('string').

#### Logical Operators, Comparison Operators and Comparators

Filters can be combined together with logical operators. Some filters take a combination of comparison operators and comparators. Following is the list of each.

##### Logical Operators

- AND - the key-value must pass both the filters to be included in the results.
- OR - the key-value must pass at least one of the filters to be included in the results.
- SKIP - for a particular row, if any of the key-values do not pass the filter condition, the entire row is skipped.
- WHILE - For a particular row, it continues to emit key-values until a key-value is reached that fails the filter condition.
- Compound Filters - Using these operators, a hierarchy of filters can be created. For example:

```
(Filter1 AND Filter2)OR(Filter3 AND Filter4)
```

##### Comparison Operators

- LESS (<)
- LESS\_OR\_EQUAL (<=)



- EQUAL (=)
- NOT\_EQUAL (!=)
- GREATER\_OR\_EQUAL (>=)
- GREATER (>)
- NO\_OP (no operation)

### Comparators

- **BinaryComparator** - lexicographically compares against the specified byte array using the `Bytes.compareTo(byte[], byte[])` method.
- **BinaryPrefixComparator** - lexicographically compares against a specified byte array. It only compares up to the length of this byte array.
- **RegexStringComparator** - compares against the specified byte array using the given regular expression. Only `EQUAL` and `NOT_EQUAL` comparisons are valid with this comparator.
- **SubStringComparator** - tests whether or not the given substring appears in a specified byte array. The comparison is case insensitive. Only `EQUAL` and `NOT_EQUAL` comparisons are valid with this comparator.

### Examples

```
Example1: >, 'binary:abc' will match everything that is lexicographically greater than
"abc"
Example2: =, 'binaryprefix:abc' will match everything whose first 3 characters are
lexicographically equal to "abc"
Example3: !=, 'regexstring:ab*yz' will match everything that doesn't begin with "ab"
and ends with "yz"
Example4: =, 'substring:abc123' will match everything that begins with the substring
"abc123"
```

### Compound Operators

Within an expression, parentheses can be used to group clauses together, and parentheses have the highest order of precedence.

`SKIP` and `WHILE` operators are next, and have the same precedence.

The `AND` operator is next.

The `OR` operator is next.

### Examples

```
A filter string of the form: "Filter1 AND Filter2 OR Filter3" will be evaluated as:
"(Filter1 AND Filter2) OR Filter3"
A filter string of the form: "Filter1 AND SKIP Filter2 OR Filter3" will be evaluated
as: "(Filter1 AND (SKIP Filter2)) OR Filter3"
```

### Filter Types

HBase includes several filter types, as well as the ability to group filters together and create your own custom filters.

- **KeyOnlyFilter** - takes no arguments. Returns the key portion of each key-value pair.

```
Syntax: KeyOnlyFilter ()
```

- **FirstKeyOnlyFilter** - takes no arguments. Returns the key portion of the first key-value pair.

```
Syntax: FirstKeyOnlyFilter ()
```

- **PrefixFilter** - takes a single argument, a prefix of a row key. It returns only those key-values present in a row that start with the specified row prefix

```
Syntax: PrefixFilter (<row_prefix>)  
Example: PrefixFilter ('Row')
```

- **ColumnPrefixFilter** - takes a single argument, a column prefix. It returns only those key-values present in a column that starts with the specified column prefix.

```
Syntax: ColumnPrefixFilter (<column_prefix>)  
Example: ColumnPrefixFilter ('Col')
```

- **MultipleColumnPrefixFilter** - takes a list of column prefixes. It returns key-values that are present in a column that starts with *any* of the specified column prefixes.

```
Syntax: MultipleColumnPrefixFilter (<column_prefix>, <column_prefix>, ..., <column_prefix>)  
Example: MultipleColumnPrefixFilter ('Col1', 'Col2')
```

- **ColumnCountGetFilter** - takes one argument, a limit. It returns the first `limit` number of columns in the table.

```
Syntax: ColumnCountGetFilter (<limit>)  
Example: ColumnCountGetFilter (4)
```

- **PageFilter** - takes one argument, a page size. It returns `page_size` number of rows from the table.

```
Syntax: PageFilter (<page_size>)  
Example: PageFilter (2)
```

- **ColumnPaginationFilter** - takes two arguments, a limit and offset. It returns limit number of columns after offset number of columns. It does this for all the rows.

```
Syntax: ColumnPaginationFilter (<limit>, <offset>)  
Example: ColumnPaginationFilter (3, 5)
```

- **InclusiveStopFilter** - takes one argument, a row key on which to stop scanning. It returns all key-values present in rows *up to and including* the specified row.

```
Syntax: InclusiveStopFilter (<stop_row_key>)  
Example: InclusiveStopFilter ('Row2')
```

- **TimeStampsFilter** - takes a list of timestamps. It returns those key-values whose timestamps matches *any* of the specified timestamps.

```
Syntax: TimeStampsFilter (<timestamp>, <timestamp>, ... ,<timestamp>)  
Example: TimeStampsFilter (5985489, 48895495, 58489845945)
```

- **RowFilter** - takes a compare operator and a comparator. It compares each row key with the comparator using the compare operator and if the comparison returns `true`, it returns all the key-values in that row.

Syntax: `RowFilter (<compareOp>, '<row_comparator>')`

Example: `RowFilter (<=, 'binary:xyz')`

- **FamilyFilter** - takes a compare operator and a comparator. It compares each family name with the comparator using the compare operator and if the comparison returns `true`, it returns all the key-values in that family.

Syntax: `FamilyFilter (<compareOp>, '<family_comparator>')`

Example: `FamilyFilter (>=, 'binaryprefix:FamilyB')`

- **QualifierFilter** - takes a compare operator and a comparator. It compares each qualifier name with the comparator using the compare operator and if the comparison returns `true`, it returns all the key-values in that column.

Syntax: `QualifierFilter (<compareOp>, '<qualifier_comparator>')`

Example: `QualifierFilter (=, 'substring:Column1')`

- **ValueFilter** - takes a compare operator and a comparator. It compares each value with the comparator using the compare operator and if the comparison returns `true`, it returns that key-value.

Syntax: `ValueFilter (<compareOp>, '<value_comparator>')`

Example: `ValueFilter (!=, 'binary:Value')`

- **DependentColumnFilter** - takes two arguments required arguments, a family and a qualifier. It tries to locate this column in each row and returns all key-values in that row that have the same timestamp. If the row does not contain the specified column, none of the key-values in that row will be returned.

The filter can also take an optional boolean argument, `dropDependentColumn`. If set to `true`, the column used for the filter does not get returned.

The filter can also take two more additional optional arguments, a compare operator and a value comparator, which are further checks in addition to the family and qualifier. If the dependent column is found, its value should also pass the value check. If it does pass the value check, only then is its timestamp taken into consideration.

Syntax: `DependentColumnFilter ('<family>', '<qualifier>', <boolean>, <compare operator>, '<value comparator>')`

`DependentColumnFilter ('<family>', '<qualifier>', <boolean>)`

`DependentColumnFilter ('<family>', '<qualifier>')`

Example: `DependentColumnFilter ('conf', 'blacklist', false, >=, 'zebra')`

`DependentColumnFilter ('conf', 'blacklist', true)`

`DependentColumnFilter ('conf', 'blacklist')`

- **SingleColumnValueFilter** - takes a column family, a qualifier, a compare operator and a comparator. If the specified column is not found, all the columns of that row will be emitted. If the column is found and the comparison with the comparator returns `true`, all the columns of the row will be emitted. If the condition fails, the row will not be emitted.

This filter also takes two additional optional boolean arguments, `filterIfColumnMissing` and `setLatestVersionOnly`.

If the `filterIfColumnMissing` flag is set to `true`, the columns of the row will not be emitted if the specified column to check is not found in the row. The default value is `false`.

If the `setLatestVersionOnly` flag is set to `false`, it will test previous versions (timestamps) in addition to the most recent. The default value is `true`.

These flags are optional and dependent on each other. You must set neither or both of them together.

```
Syntax: SingleColumnValueFilter ('<family>', '<qualifier>', <compare operator>,
'<comparator>', <filterIfColumnMissing_boolean>, <latest_version_boolean>)
```

```
Syntax: SingleColumnValueFilter ('<family>', '<qualifier>', <compare operator>,
'<comparator>')
```

```
Example: SingleColumnValueFilter ('FamilyA', 'Column1', <=, 'abc', true, false)
```

```
Example: SingleColumnValueFilter ('FamilyA', 'Column1', <=, 'abc')
```

- **SingleColumnValueExcludeFilter** - takes the same arguments and behaves same as `SingleColumnValueFilter`. However, if the column is found and the condition passes, all the columns of the row will be emitted except for the tested column value.

```
Syntax: SingleColumnValueExcludeFilter (<family>, <qualifier>, <compare operators>,
<comparator>, <latest_version_boolean>, <filterIfColumnMissing_boolean>)
```

```
Syntax: SingleColumnValueExcludeFilter (<family>, <qualifier>, <compare operator>
<comparator>)
```

```
Example: SingleColumnValueExcludeFilter ('FamilyA', 'Column1', '<=', 'abc', 'false',
'true')
```

```
Example: SingleColumnValueExcludeFilter ('FamilyA', 'Column1', '<=', 'abc')
```

- **ColumnRangeFilter** - takes either `minColumn`, `maxColumn`, or both. Returns only those keys with columns that are between `minColumn` and `maxColumn`. It also takes two boolean variables to indicate whether to include the `minColumn` and `maxColumn` or not. If you don't want to set the `minColumn` or the `maxColumn`, you can pass in an empty argument.

```
Syntax: ColumnRangeFilter ('<minColumn >', <minColumnInclusive_bool>,
'<maxColumn>', <maxColumnInclusive_bool>)
```

```
Example: ColumnRangeFilter ('abc', true, 'xyz', false)
```

- **Custom Filter** - You can create a custom filter by implementing the [Filter](#) class. The JAR must be available on all region servers.

### HBase Shell Example

This example scans the 'users' table for rows where the contents of the `cf:name` column equals the string 'abc'.

```
hbase> scan 'users', { FILTER => SingleColumnValueFilter.new(Bytes.toBytes('cf'),
Bytes.toBytes('name'), CompareFilter::CompareOp.valueOf('EQUAL'),
BinaryComparator.new(Bytes.toBytes('abc')))}
```

### Java API Example

This example, taken from the HBase unit test found in

`hbase-server/src/test/java/org/apache/hadoop/hbase/filter/TestSingleColumnValueFilter.java`, shows how to use the Java API to implement several different filters..

```
/**
 *
 * Licensed to the Apache Software Foundation (ASF) under one
 * or more contributor license agreements. See the NOTICE file
 * distributed with this work for additional information
 * regarding copyright ownership. The ASF licenses this file
 * to you under the Apache License, Version 2.0 (the
```

```

* "License"); you may not use this file except in compliance
* with the License. You may obtain a copy of the License at
*
*     http://www.apache.org/licenses/LICENSE-2.0
*
* Unless required by applicable law or agreed to in writing, software
* distributed under the License is distributed on an "AS IS" BASIS,
* WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
* See the License for the specific language governing permissions and
* limitations under the License.
*/
package org.apache.hadoop.hbase.filter;

import static org.junit.Assert.assertFalse;
import static org.junit.Assert.assertTrue;

import java.util.regex.Pattern;

import org.apache.hadoop.hbase.KeyValue;
import org.apache.hadoop.hbase.SmallTests;
import org.apache.hadoop.hbase.filter.CompareFilter.CompareOp;
import org.apache.hadoop.hbase.util.Bytes;
import org.junit.Before;
import org.junit.Test;
import org.junit.experimental.categories.Category;

/**
 * Tests the value filter
 */
@Category(SmallTests.class)
public class TestSingleColumnValueFilter {
    private static final byte[] ROW = Bytes.toBytes("test");
    private static final byte[] COLUMN_FAMILY = Bytes.toBytes("test");
    private static final byte[] COLUMN_QUALIFIER = Bytes.toBytes("foo");
    private static final byte[] VAL_1 = Bytes.toBytes("a");
    private static final byte[] VAL_2 = Bytes.toBytes("ab");
    private static final byte[] VAL_3 = Bytes.toBytes("abc");
    private static final byte[] VAL_4 = Bytes.toBytes("abcd");
    private static final byte[] FULLSTRING_1 =
        Bytes.toBytes("The quick brown fox jumps over the lazy dog.");
    private static final byte[] FULLSTRING_2 =
        Bytes.toBytes("The slow grey fox trips over the lazy dog.");
    private static final String QUICK_SUBSTR = "quick";
    private static final String QUICK_REGEX = ".+quick.+";
    private static final Pattern QUICK_PATTERN = Pattern.compile("QuIcK",
        Pattern.CASE_INSENSITIVE | Pattern.DOTALL);

    Filter basicFilter;
    Filter nullFilter;
    Filter substrFilter;
    Filter regexFilter;
    Filter regexPatternFilter;

    @Before
    public void setUp() throws Exception {
        basicFilter = basicFilterNew();
        nullFilter = nullFilterNew();
        substrFilter = substrFilterNew();
        regexFilter = regexFilterNew();
        regexPatternFilter = regexFilterNew(QUICK_PATTERN);
    }

    private Filter basicFilterNew() {
        return new SingleColumnValueFilter(COLUMN_FAMILY, COLUMN_QUALIFIER,
            CompareOp.GREATER_OR_EQUAL, VAL_2);
    }

    private Filter nullFilterNew() {
        return new SingleColumnValueFilter(COLUMN_FAMILY, COLUMN_QUALIFIER,
            CompareOp.NOT_EQUAL,
                new NullComparator());
    }
}

```

```

private Filter substrFilterNew() {
    return new SingleColumnValueFilter(COLUMN_FAMILY, COLUMN_QUALIFIER,
        CompareOp.EQUAL,
        new SubstringComparator(QUICK_SUBSTR));
}

private Filter regexFilterNew() {
    return new SingleColumnValueFilter(COLUMN_FAMILY, COLUMN_QUALIFIER,
        CompareOp.EQUAL,
        new RegexStringComparator(QUICK_REGEX));
}

private Filter regexFilterNew(Pattern pattern) {
    return new SingleColumnValueFilter(COLUMN_FAMILY, COLUMN_QUALIFIER,
        CompareOp.EQUAL,
        new RegexStringComparator(pattern.pattern(), pattern.flags()));
}

private void basicFilterTests(SingleColumnValueFilter filter)
    throws Exception {
    KeyValue kv = new KeyValue(ROW, COLUMN_FAMILY, COLUMN_QUALIFIER, VAL_2);
    assertTrue("basicFilter1", filter.filterKeyValue(kv) == Filter.ReturnCode.INCLUDE);

    kv = new KeyValue(ROW, COLUMN_FAMILY, COLUMN_QUALIFIER, VAL_3);
    assertTrue("basicFilter2", filter.filterKeyValue(kv) == Filter.ReturnCode.INCLUDE);

    kv = new KeyValue(ROW, COLUMN_FAMILY, COLUMN_QUALIFIER, VAL_4);
    assertTrue("basicFilter3", filter.filterKeyValue(kv) == Filter.ReturnCode.INCLUDE);

    assertFalse("basicFilterNotNull", filter.filterRow());
    filter.reset();
    kv = new KeyValue(ROW, COLUMN_FAMILY, COLUMN_QUALIFIER, VAL_1);
    assertTrue("basicFilter4", filter.filterKeyValue(kv) == Filter.ReturnCode.NEXT_ROW);

    kv = new KeyValue(ROW, COLUMN_FAMILY, COLUMN_QUALIFIER, VAL_2);
    assertTrue("basicFilter4", filter.filterKeyValue(kv) == Filter.ReturnCode.NEXT_ROW);

    assertFalse("basicFilterAllRemaining", filter.filterAllRemaining());
    assertTrue("basicFilterNotNull", filter.filterRow());
    filter.reset();
    filter.setLatestVersionOnly(false);
    kv = new KeyValue(ROW, COLUMN_FAMILY, COLUMN_QUALIFIER, VAL_1);
    assertTrue("basicFilter5", filter.filterKeyValue(kv) == Filter.ReturnCode.INCLUDE);

    kv = new KeyValue(ROW, COLUMN_FAMILY, COLUMN_QUALIFIER, VAL_2);
    assertTrue("basicFilter5", filter.filterKeyValue(kv) == Filter.ReturnCode.INCLUDE);

    assertFalse("basicFilterNotNull", filter.filterRow());
}

private void nullFilterTests(Filter filter) throws Exception {
    ((SingleColumnValueFilter) filter).setFilterIfMissing(true);
    KeyValue kv = new KeyValue(ROW, COLUMN_FAMILY, COLUMN_QUALIFIER, FULLSTRING_1);
    assertTrue("null1", filter.filterKeyValue(kv) == Filter.ReturnCode.INCLUDE);
    assertFalse("nullFilterRow", filter.filterRow());
    filter.reset();
    kv = new KeyValue(ROW, COLUMN_FAMILY, Bytes.toBytes("qual2"), FULLSTRING_2);
    assertTrue("null2", filter.filterKeyValue(kv) == Filter.ReturnCode.INCLUDE);
    assertTrue("null2FilterRow", filter.filterRow());
}

private void substrFilterTests(Filter filter)
    throws Exception {
    KeyValue kv = new KeyValue(ROW, COLUMN_FAMILY, COLUMN_QUALIFIER,
        FULLSTRING_1);
    assertTrue("substrTrue",
        filter.filterKeyValue(kv) == Filter.ReturnCode.INCLUDE);
    kv = new KeyValue(ROW, COLUMN_FAMILY, COLUMN_QUALIFIER,
        FULLSTRING_2);
    assertTrue("substrFalse", filter.filterKeyValue(kv) == Filter.ReturnCode.INCLUDE);

    assertFalse("substrFilterAllRemaining", filter.filterAllRemaining());
    assertFalse("substrFilterNotNull", filter.filterRow());
}

```

```

    }

    private void regexFilterTests(Filter filter)
        throws Exception {
        KeyValue kv = new KeyValue(ROW, COLUMN_FAMILY, COLUMN_QUALIFIER,
            FULLSTRING_1);
        assertTrue("regexTrue",
            filter.filterKeyValue(kv) == Filter.ReturnCode.INCLUDE);
        kv = new KeyValue(ROW, COLUMN_FAMILY, COLUMN_QUALIFIER,
            FULLSTRING_2);
        assertTrue("regexFalse", filter.filterKeyValue(kv) == Filter.ReturnCode.INCLUDE);
        assertFalse("regexFilterAllRemaining", filter.filterAllRemaining());
        assertFalse("regexFilterNotNull", filter.filterRow());
    }

    private void regexPatternFilterTests(Filter filter)
        throws Exception {
        KeyValue kv = new KeyValue(ROW, COLUMN_FAMILY, COLUMN_QUALIFIER,
            FULLSTRING_1);
        assertTrue("regexTrue",
            filter.filterKeyValue(kv) == Filter.ReturnCode.INCLUDE);
        assertFalse("regexFilterAllRemaining", filter.filterAllRemaining());
        assertFalse("regexFilterNotNull", filter.filterRow());
    }

    private Filter serializationTest(Filter filter)
        throws Exception {
        // Decompose filter to bytes.
        byte[] buffer = filter.toByteArray();

        // Recompose filter.
        Filter newFilter = SingleColumnValueFilter.parseFrom(buffer);
        return newFilter;
    }

    /**
     * Tests identification of the stop row
     * @throws Exception
     */
    @Test
    public void testStop() throws Exception {
        basicFilterTests((SingleColumnValueFilter) basicFilter);
        nullFilterTests(nullFilter);
        substrFilterTests(substrFilter);
        regexFilterTests(regexFilter);
        regexPatternFilterTests(regexPatternFilter);
    }

    /**
     * Tests serialization
     * @throws Exception
     */
    @Test
    public void testSerialization() throws Exception {
        Filter newFilter = serializationTest(basicFilter);
        basicFilterTests((SingleColumnValueFilter) newFilter);
        newFilter = serializationTest(nullFilter);
        nullFilterTests(newFilter);
        newFilter = serializationTest(substrFilter);
        substrFilterTests(newFilter);
        newFilter = serializationTest(regexFilter);
        regexFilterTests(newFilter);
        newFilter = serializationTest(regexPatternFilter);
        regexPatternFilterTests(newFilter);
    }
}

```

## Writing Data to HBase

To write data to HBase, you use methods of the `HTableInterface` class. You can use the Java API directly, or use HBase Shell, Thrift API, REST API, or another client which uses the Java API indirectly. When you issue a Put,

the coordinates of the data are the row, the column, and the timestamp. The timestamp is unique per version of the cell, and can be generated automatically or specified programmatically by your application, and must be a long integer.

### Variations on Put

There are several different ways to write data into HBase. Some of them are listed below.

- A `Put` operation writes data into HBase.
- A `Delete` operation deletes data from HBase. What actually happens during a `Delete` depends upon several factors.
- A `CheckAndPut` operation performs a `Scan` before attempting the `Put`, and only does the `Put` if a value matches what is expected, and provides row-level atomicity.
- A `CheckAndDelete` operation performs a `Scan` before attempting the `Delete`, and only does the `Delete` if a value matches what is expected.
- An `Increment` operation increments values of one or more columns within a single row, and provides row-level atomicity.

Refer to the API documentation for a full list of methods provided for writing data to HBase. Different methods require different access levels and have other differences.

### Versions

When you put data into HBase, a timestamp is required. The timestamp can be generated automatically by the `RegionServer` or can be supplied by you. The timestamp must be unique per version of a given cell, because the timestamp identifies the version. To modify a previous version of a cell, for instance, you would issue a `Put` with a different value for the data itself, but the same timestamp.

HBase's behavior regarding versions is highly configurable. The maximum number of versions defaults to 1 in CDH 5, and 3 in previous versions. You can change the default value for HBase by configuring `hbase.column.max.version` in `hbase-site.xml`, either via an advanced configuration snippet if you use Cloudera Manager, or by editing the file directly otherwise.

You can also configure the maximum and minimum number of versions to keep for a given column, or specify a default time-to-live (TTL), which is the number of seconds before a version is deleted. The following examples all use `alter` statements in HBase Shell to create new column families with the given characteristics, but you can use the same syntax when creating a new table or to alter an existing column family. This is only a fraction of the options you can specify for a given column family.

```
hbase> alter 't1', NAME => 'f1', VERSIONS => 5
hbase> alter 't1', NAME => 'f1', MIN_VERSIONS => 2
hbase> alter 't1', NAME => 'f1', TTL => 15
```

HBase sorts the versions of a cell from newest to oldest, by sorting the timestamps lexicographically. When a version needs to be deleted because a threshold has been reached, HBase always chooses the "oldest" version, even if it is in fact the most recent version to be inserted. Keep this in mind when designing your timestamps. Consider using the default generated timestamps and storing other version-specific data elsewhere in the row, such as in the row key. If `MIN_VERSIONS` and `TTL` conflict, `MIN_VERSIONS` takes precedence.

### Deletion

When you request for HBase to delete data, either explicitly via a `Delete` method or implicitly via a threshold such as the maximum number of versions or the TTL, HBase does not delete the data immediately. Instead, it writes a deletion marker, called a tombstone, to the HFile, which is the physical file where a given `RegionServer` stores its region of a column family. The tombstone markers are processed during major compaction operations, when HFiles are rewritten without the deleted data included.



Even after major compactions, "deleted" data may not actually be deleted. You can specify the `KEEP_DELETED_CELLS` option for a given column family, and the tombstones will be preserved in the HFile even after major compaction. One scenario where this approach might be useful is for data retention policies.

Another reason deleted data may not actually be deleted is if the data would be required to restore a table from a snapshot which has not been deleted. In this case, the data is moved to an archive during a major compaction, and only deleted when the snapshot is deleted. This is a good reason to monitor the number of snapshots saved in HBase.

## Examples

This abbreviated example writes data to an HBase table using HBase Shell and then scans the table to show the result.

```
hbase> put 'test', 'row1', 'cf:a', 'value1'
0 row(s) in 0.1770 seconds

hbase> put 'test', 'row2', 'cf:b', 'value2'
0 row(s) in 0.0160 seconds

hbase> put 'test', 'row3', 'cf:c', 'value3'
0 row(s) in 0.0260 seconds
hbase> scan 'test'
ROW                                COLUMN+CELL
 row1                             column=cf:a, timestamp=1403759475114, value=value1
 row2                             column=cf:b, timestamp=1403759492807, value=value2
 row3                             column=cf:c, timestamp=1403759503155, value=value3
3 row(s) in 0.0440 seconds
```

This abbreviated example uses the HBase API to write data to an HBase table, using the automatic timestamp created by the Region Server.

```
publicstaticfinalbyte[] CF = "cf".getBytes();
publicstaticfinalbyte[] ATTR = "attr".getBytes();
...
Put put = new Put(Bytes.toBytes(row));
put.add(CF, ATTR, Bytes.toBytes(data));
htable.put(put);
```

This example uses the HBase API to write data to an HBase table, specifying the timestamp.

```
publicstaticfinalbyte[] CF = "cf".getBytes();
publicstaticfinalbyte[] ATTR = "attr".getBytes();
...
Put put = new Put(Bytes.toBytes(row));
long explicitTimeInMs = 555; // just an example
put.add(CF, ATTR, explicitTimeInMs, Bytes.toBytes(data));
htable.put(put);
```

## Further Reading

- Refer to the [HTableInterface](#) and [HColumnDescriptor](#) API documentation for more details about configuring tables and columns, as well as reading and writing to HBase.
- Refer to the [Apache HBase Reference Guide](#) for more in-depth information about HBase, including details about versions and deletions not covered here.

## Importing Data Into HBase

The method you use for importing data into HBase depends on several factors:

- The location, size, and format of your existing data
- Whether you need to import data once or periodically over time
- Whether you want to import the data in bulk or stream it into HBase regularly
- How fresh the HBase data needs to be

This topic helps you choose the correct method or composite of methods and provides example workflows for each method.

### Choosing the Right Import Method

#### If the data is already in an HBase table:

- To move the data from one HBase cluster to another, use `snapshot` and either the `clone_snapshot` or `ExportSnapshot` utility; or, use the `CopyTable` utility.
- To move the data from one HBase cluster to another without downtime on either cluster, use replication.

#### If the data currently exists outside HBase:

- If possible, write the data to HFile format, and use a `BulkLoad` to import it into HBase. The data is immediately available to HBase and you can bypass the normal write path, increasing efficiency.
- If you prefer not to use bulk loads, and you are using a tool such as Pig, you can use it to import your data.

#### If you need to stream live data to HBase instead of import in bulk:

- Write a Java client using the Java API, or use the Apache Thrift Proxy API to write a client in a language supported by Thrift.
- Stream data directly into HBase using the REST Proxy API in conjunction with an HTTP client such as `wget` or `curl`.
- Use Flume or Spark.

Most likely, at least one of these methods works in your situation. If not, you can use MapReduce directly. Test the most feasible methods with a subset of your data to determine which one is optimal.

### Using CopyTable

`CopyTable` uses HBase read and write paths to copy part or all of a table to a new table in either the same cluster or a different cluster. `CopyTable` causes read load when reading from the source, and write load when writing to the destination. Region splits occur on the destination table in real time as needed. To avoid these issues, use `snapshot` and `export` commands instead of `CopyTable`. Alternatively, you can pre-split the destination table to avoid excessive splits. The destination table can be partitioned differently from the source table. See [this section](#) of the Apache HBase documentation for more information.

Edits to the source table after the `CopyTable` starts are not copied, so you may need to do an additional `CopyTable` operation to copy new data into the destination table. Run `CopyTable` as follows, using `--help` to see details about possible parameters.

```
$ ./bin/hbase org.apache.hadoop.hbase.mapreduce.CopyTable --help
Usage: CopyTable [general options] [--starttime=X] [--endtime=Y] [--new.name=NEW]
[--peer.adr=ADR] <tablename>
```

The `starttime/endtime` and `startrow/endrow` pairs function in a similar way: if you leave out the first of the pair, the first timestamp or row in the table is the starting point. Similarly, if you leave out the second of the pair, the operation continues until the end of the table. To copy the table to a new table in the same cluster, you must specify `--new.name`, unless you want to write the copy back to the same table, which would add a new version of each cell (with the same data), or just overwrite the cell with the same value if the maximum number of versions is set to 1 (the default in CDH 5). To copy the table to a new table in a different cluster, specify `--peer.adr` and optionally, specify a new table name.

The following example creates a new table using HBase Shell in non-interactive mode, and then copies data in two ColumnFamilies in rows starting with timestamp 1265875194289 and including the last row before the CopyTable started, to the new table.

```
$ echo create 'NewTestTable', 'cf1', 'cf2', 'cf3' | bin/hbase shell --non-interactive
$ bin/hbase org.apache.hadoop.hbase.mapreduce.CopyTable --starttime=1265875194289
--families=cf1,cf2,cf3 --new.name=NewTestTable TestTable
```

In CDH 5, snapshots are recommended instead of CopyTable for most situations.

### Using Snapshots

As of CDH 4.7, Cloudera recommends snapshots instead of CopyTable where possible. A snapshot captures the state of a table at the time the snapshot was taken. Because no data is copied when a snapshot is taken, the process is very quick. As long as the snapshot exists, cells in the snapshot are never deleted from HBase, even if they are explicitly deleted by the API. Instead, they are archived so that the snapshot can restore the table to its state at the time of the snapshot.

After taking a snapshot, use the `clone_snapshot` command to copy the data to a new (immediately enabled) table in the same cluster, or the Export utility to create a new table based on the snapshot, in the same cluster or a new cluster. This is a copy-on-write operation. The new table shares HFiles with the original table until writes occur in the new table but not the old table, or until a compaction or split occurs in either of the tables. This can improve performance in the short term compared to CopyTable.

To export the snapshot to a new cluster, use the `ExportSnapshot` utility, which uses MapReduce to copy the snapshot to the new cluster. Run the `ExportSnapshot` utility on the source cluster, as a user with HBase and HDFS write permission on the destination cluster, and HDFS read permission on the source cluster. This creates the expected amount of IO load on the destination cluster. Optionally, you can limit bandwidth consumption, which affects IO on the destination cluster. After the `ExportSnapshot` operation completes, you can see the snapshot in the new cluster using the `list_snapshot` command, and you can use the `clone_snapshot` command to create the table in the new cluster from the snapshot.

For full instructions for the `snapshot` and `clone_snapshot` HBase Shell commands, run the HBase Shell and type `help snapshot`. The following example takes a snapshot of a table, uses it to clone the table to a new table in the same cluster, and then uses the `ExportSnapshot` utility to copy the table to a different cluster, with 16 mappers and limited to 200 Mb/sec bandwidth.

```
$ bin/hbase shell
hbase(main):005:0> snapshot 'TestTable', 'TestTableSnapshot'
0 row(s) in 2.3290 seconds

hbase(main):006:0> clone_snapshot 'TestTableSnapshot', 'NewTestTable'
0 row(s) in 1.3270 seconds

hbase(main):007:0> describe 'NewTestTable'
DESCRIPTION                                     ENABLED
'NewTestTable', {NAME => 'cf1', DATA_BLOCK_ENCODING => 'NONE', BLOOMFILTER => 'ROW', REPLICATION_SCOPE => '0', VERSIONS => '1', COMPRESSION => 'NONE', MIN_VERSIONS => '0', TTL => 'FOREVER', KEEP_DELETED_CELLS => 'false', BLOCKSIZE => '65536', IN_MEMORY => 'false', BLOCKCACHE => 'true'}, {NAME => 'cf2', DATA_BLOCK_ENCODING => 'NONE', BLOOMFILTER => 'ROW', REPLICATION_SCOPE => '0', VERSIONS => '1', COMPRESSION => 'NONE', MIN_VERSIONS => '0', TTL => 'FOREVER', KEEP_DELETED_CELLS => 'false', BLOCKSIZE => '65536', IN_MEMORY => 'false', BLOCKCACHE => 'true'}
1 row(s) in 0.1280 seconds
hbase(main):008:0> quit

$ hbase org.apache.hadoop.hbase.snapshot.ExportSnapshot -snapshot TestTableSnapshot
-copy-to file:///tmp/hbase -mappers 16 -bandwidth 200
14/10/28 21:48:16 INFO snapshot.ExportSnapshot: Copy Snapshot Manifest
14/10/28 21:48:17 INFO client.RMProxy: Connecting to ResourceManager at
a1221.halxg.cloudera.com/10.20.188.121:8032
14/10/28 21:48:19 INFO snapshot.ExportSnapshot: Loading Snapshot 'TestTableSnapshot'
```

```
hfile list
14/10/28 21:48:19 INFO Configuration.deprecation: hadoop.native.lib is deprecated.
Instead, use io.native.lib.available
14/10/28 21:48:19 INFO util.FSVisitor: No logs under
directory:hdfs://a1221.halxg.cloudera.com:8020/hbase/.hbase-snapshot/TestTableSnapshot/WALs
14/10/28 21:48:20 INFO mapreduce.JobSubmitter: number of splits:0
14/10/28 21:48:20 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_1414556809048_0001
14/10/28 21:48:20 INFO impl.YarnClientImpl: Submitted application
application_1414556809048_0001
14/10/28 21:48:20 INFO mapreduce.Job: The url to track the job:
http://a1221.halxg.cloudera.com:8088/proxy/application_1414556809048_0001/
14/10/28 21:48:20 INFO mapreduce.Job: Running job: job_1414556809048_0001
14/10/28 21:48:36 INFO mapreduce.Job: Job job_1414556809048_0001 running in uber mode
: false
14/10/28 21:48:36 INFO mapreduce.Job: map 0% reduce 0%
14/10/28 21:48:37 INFO mapreduce.Job: Job job_1414556809048_0001 completed successfully
14/10/28 21:48:37 INFO mapreduce.Job: Counters: 2
Job Counters
  Total time spent by all maps in occupied slots (ms)=0
  Total time spent by all reduces in occupied slots (ms)=0
14/10/28 21:48:37 INFO snapshot.ExportSnapshot: Finalize the Snapshot Export
14/10/28 21:48:37 INFO snapshot.ExportSnapshot: Verify snapshot integrity
14/10/28 21:48:37 INFO Configuration.deprecation: fs.default.name is deprecated. Instead,
use fs.defaultFS
14/10/28 21:48:37 INFO snapshot.ExportSnapshot: Export Completed: TestTableSnapshot
```

The bold italic line contains the URL from which you can track the `ExportSnapshot` job. When it finishes, a new set of HFiles, comprising all of the HFiles that were part of the table when the snapshot was taken, is created at the HDFS location you specified.

You can use the `SnapshotInfo` command-line utility included with HBase to verify or debug snapshots.

### Using BulkLoad

HBase uses the well-known HFile format to store its data on disk. In many situations, writing HFiles programmatically with your data, and bulk-loading that data into HBase on the `RegionServer`, has advantages over other data ingest mechanisms. BulkLoad operations bypass the write path completely, providing the following benefits:

- The data is available to HBase immediately but does cause additional load or latency on the cluster when it appears.
- BulkLoad operations do not use the write-ahead log (WAL) and do not cause flushes or split storms.
- BulkLoad operations do not cause excessive garbage collection.

- **Note:** Because they bypass the WAL, BulkLoad operations are not propagated between clusters using replication. If you need the data on all replicated clusters, you must perform the BulkLoad on each cluster.

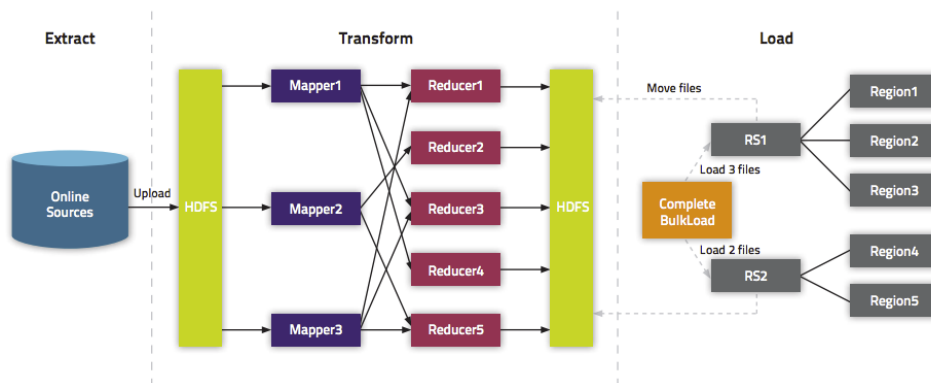
If you use BulkLoads with HBase, your workflow is similar to the following:

1. **Extract your data from its existing source.** For instance, if your data is in a MySQL database, you might run the `mysqldump` command. The process you use depends on your data. If your data is already in TSV or CSV format, skip this step and use the included `ImportTsv` utility to process your data into HFiles. See the [ImportTsv documentation](#) for details.
2. **Process your data into HFile format.** See [http://hbase.apache.org/book/hfile\\_format.html](http://hbase.apache.org/book/hfile_format.html) for details about HFile format. Usually you use a MapReduce job for the conversion, and you often need to write the Mapper yourself because your data is unique. The job must to emit the row key as the `Key`, and either a `KeyValue`, a `Put`, or a `Delete` as the `Value`. The Reducer is handled by HBase; configure it using [HFileOutputFormat.configureIncrementalLoad\(\)](#) and it does the following:
  - Inspects the table to configure a total order partitioner
  - Uploads the partitions file to the cluster and adds it to the `DistributedCache`
  - Sets the number of `reduce` tasks to match the current number of regions

- Sets the output key/value class to match `HFileOutputFormat` requirements
  - Sets the Reducer to perform the appropriate sorting (either `KeyValueSortReducer` or `PutSortReducer`)
3. **One HFile is created per region in the output folder.** Input data is almost completely re-written, so you need available disk space at least twice the size of the original data set. For example, for a 100 GB output from `mysqldump`, you should have at least 200 GB of available disk space in HDFS. You can delete the original input file at the end of the process.
  4. **Load the files into HBase.** Use the `LoadIncrementalHFiles` command (more commonly known as the [completebulkload](#) tool), passing it a URL that locates the files in HDFS. Each file is loaded into the relevant region on the RegionServer for the region. You can limit the number of versions that are loaded by passing the `--versions= N` option, where `N` is the maximum number of versions to include, from newest to oldest (largest timestamp to smallest timestamp).

If a region was split after the files were created, the tool automatically splits the HFile according to the new boundaries. This process is inefficient, so if your table is being written to by other processes, you should load as soon as the transform step is done.

The following illustration shows the full BulkLoad process.



#### Use Cases for BulkLoad:

- **Loading your original dataset into HBase for the first time** - Your initial dataset might be quite large, and bypassing the HBase write path can speed up the process considerably.
- **Incremental Load** - To load new data periodically, use BulkLoad to import it in batches at your preferred intervals. This alleviates latency problems and helps you to achieve service-level agreements (SLAs). However, one trigger for compaction is the number of HFiles on a RegionServer. Therefore, importing a large number of HFiles at frequent intervals can cause major compactons to happen more often than they otherwise would, negatively impacting performance. You can mitigate this by tuning the compaction settings such that the maximum number of HFiles that can be present without triggering a compaction is very high, and relying on other factors, such as the size of the Memstore, to trigger compactons.
- **Data needs to originate elsewhere** - If an existing system is capturing the data you want to have in HBase and needs to remain active for business reasons, you can periodically BulkLoad data from the system into HBase so that you can perform operations on it without impacting the system.

For more information and examples, as well as an explanation of the `ImportTsv` utility, which can be used to import data in text-delimited formats such as CSV, see [this post](#) on the Cloudera Blog.

#### Using Cluster Replication

If your data is already in an HBase cluster, replication is useful for getting the data into additional HBase clusters. In HBase, cluster replication refers to keeping one cluster state synchronized with that of another cluster, using the write-ahead log (WAL) of the source cluster to propagate the changes. Replication is enabled at column family granularity. Before enabling replication for a column family, create the table and all column families to be replicated, on the destination cluster.

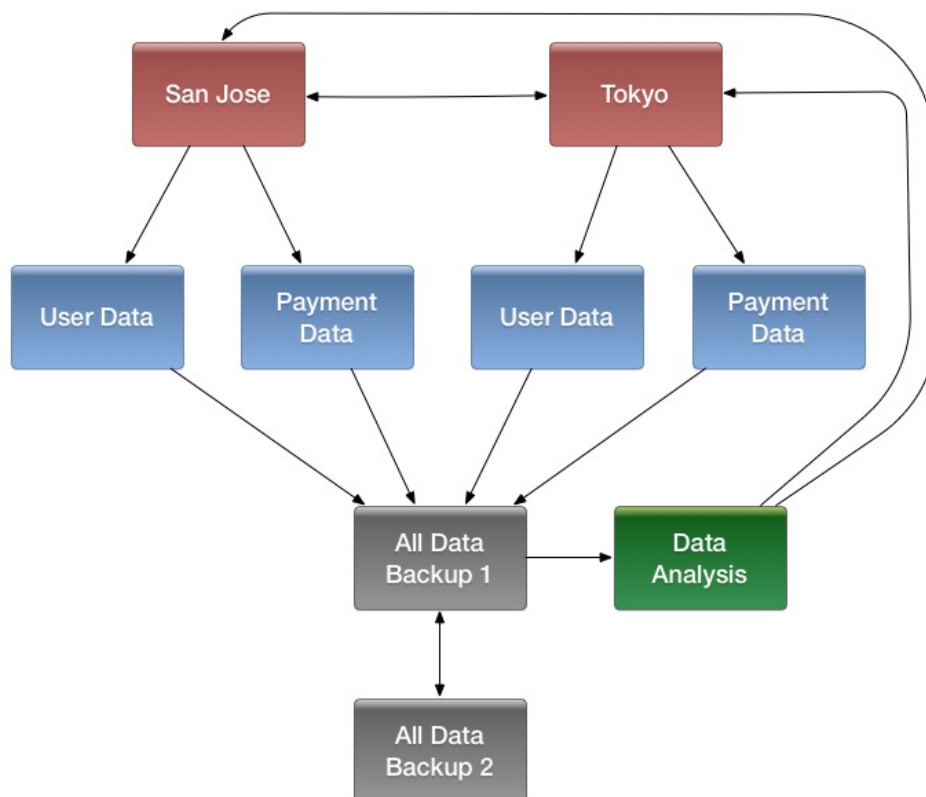
Cluster replication uses a master-push methodology. An HBase cluster can be a source (also called *master* or *active*, meaning that it is the originator of new data), a destination (also called *slave* or *passive*, meaning that it receives data via replication), or can fulfill both roles at once. Replication is asynchronous, and the goal of replication is consistency.

When data is replicated from one cluster to another, the original source of the data is tracked with a cluster ID, which is part of the metadata. In CDH 5, all clusters that have already consumed the data are also tracked. This prevents replication loops.

- **Note:** Previously, terms such as *master-master*, *master-slave*, and *cyclic* were used to describe replication relationships in HBase. These terms were confusing and have been replaced by discussions about cluster topologies appropriate for different scenarios.

### Common Replication Topologies

- A central source cluster might propagate changes to multiple destination clusters, for failover or due to geographic distribution.
- A source cluster might push changes to a destination cluster, which might also push its own changes back to the original cluster.
- Many different low-latency clusters might push changes to one centralized cluster for backup or resource-intensive data-analytics jobs. The processed data might then be replicated back to the low-latency clusters.
- Multiple levels of replication can be chained together to suit your needs. The following diagram shows a hypothetical scenario. Use the arrows to follow the data paths.



At the top of the diagram, the San Jose and Tokyo clusters, shown in red, replicate changes to each other, and each also replicates changes to a User Data and a Payment Data cluster.

Each cluster in the second row, shown in blue, replicates its changes to the All Data Backup 1 cluster, shown in grey. The All Data Backup 1 cluster replicates changes to the All Data Backup 2 cluster (also shown in

grey), as well as the Data Analysis cluster (shown in green). All Data Backup 2 also propagates any of its own changes back to All Data Backup 1.

The Data Analysis cluster runs MapReduce jobs on its data, and then pushes the processed data back to the San Jose and Tokyo clusters.

### Configuring Clusters for Replication

To configure your clusters for replication, see [HBase Replication](#) on page 269 and [Configuring Secure HBase Replication](#). The following is a high-level overview of the steps to enable replication.

1. Configure and start the source and destination clusters. Create tables with the same names and column families on both the source and destination clusters, so that the destination cluster knows where to store data it receives. All hosts in the source and destination clusters should be reachable to each other.
2. On the source cluster, enable replication in Cloudera Manager, or by setting `hbase.replication` to `true` in `hbase-site.xml`.
3. On the source cluster, in HBase Shell, add the destination cluster as a peer, using the `add_peer` command. The syntax is as follows:

```
add_peer 'ID' 'CLUSTER_KEY'
```

The ID must be a short integer. To compose the `CLUSTER_KEY`, use the following template:

```
hbase.zookeeper.quorum:hbase.zookeeper.property.clientPort:zookeeper.znode.parent
```

If both clusters use the same ZooKeeper cluster, you must use a different **zookeeper.znode.parent**, because they cannot write in the same folder.

4. On the source cluster, configure each column family to be replicated by setting its `REPLICATION_SCOPE` to 1, using commands such as the following in HBase Shell.

```
hbase> disable 'example_table'
hbase> alter 'example_table', {NAME => 'example_family', REPLICATION_SCOPE => '1'}
hbase> enable 'example_table'
```

5. Verify that replication is occurring by examining the logs on the source cluster for messages such as the following.

```
Considering 1 rs, with ratio 0.1
Getting 1 rs from peer cluster # 0
Choosing peer 10.10.1.49:62020
```

6. To verify the validity of replicated data, use the included `VerifyReplication` MapReduce job on the source cluster, providing it with the ID of the replication peer and table name to verify. Other options are available, such as a time range or specific families to verify.

The command has the following form:

```
hbase org.apache.hadoop.hbase.mapreduce.replication.VerifyReplication
[--starttime=timestamp] [--stoptime=timestamp] [--families=comma separated list
of families] <peerId> <tablename>
```

The `VerifyReplication` command prints `GOODROWS` and `BADROWS` counters to indicate rows that did and did not replicate correctly.



■ **Note:**

Some changes are not replicated and must be propagated by other means, such as [Snapshots](#) or [CopyTable](#). See [Initiating Replication When Data Already Exists](#) on page 273 for more details.

- Data that existed in the master before replication was enabled.
- Operations that bypass the WAL, such as when using BulkLoad or API calls such as `writeToWal(false)`.
- Table schema modifications.

### Using Pig and HCatalog

Apache Pig is a platform for analyzing large data sets using a high-level language. Apache HCatalog is a sub-project of Apache Hive, which enables reading and writing of data from one Hadoop utility to another. You can use a combination of Pig and HCatalog to import data into HBase. The initial format of your data and other details about your infrastructure determine the steps you follow to accomplish this task. The following simple example assumes that you can get your data into a TSV (text-separated value) format, such as a tab-delimited or comma-delimited text file.

1. Format the data as a TSV file. You can work with other file formats; see the Pig and HCatalog project documentation for more details.

The following example shows a subset of data from [Google's NGram Dataset](#), which shows the frequency of specific phrases or letter-groupings found in publications indexed by Google. Here, the first column has been added to this dataset as the row ID. It is formulated by combining the n-gram itself (in this case, `Zones`) with the line number of the file in which it is found (`z_LINE_NUM`). This creates a format such as `"Zones_z_6230867."` The second column is the n-gram itself, the third column is the year of occurrence, the fourth column is the frequency of occurrence of that Ngram in that year, and the fifth column is the number of distinct publications. This extract is from the z file of the 1-gram dataset from version 20120701. The data is truncated at the . . . mark, for the sake of readability of this document. In most real-world scenarios, you will not work with tables that have five columns. Most HBase tables have one or two columns.

```
Zones_z_6230867 Zones 1507 1 1
Zones_z_6230868 Zones 1638 1 1
Zones_z_6230869 Zones 1656 2 1
Zones_z_6230870 Zones 1681 8 2
...
Zones_z_6231150 Zones 1996 17868 4356
Zones_z_6231151 Zones 1997 21296 4675
Zones_z_6231152 Zones 1998 20365 4972
Zones_z_6231153 Zones 1999 20288 5021
Zones_z_6231154 Zones 2000 22996 5714
Zones_z_6231155 Zones 2001 20469 5470
Zones_z_6231156 Zones 2002 21338 5946
Zones_z_6231157 Zones 2003 29724 6446
Zones_z_6231158 Zones 2004 23334 6524
Zones_z_6231159 Zones 2005 24300 6580
Zones_z_6231160 Zones 2006 22362 6707
Zones_z_6231161 Zones 2007 22101 6798
Zones_z_6231162 Zones 2008 21037 6328
```

2. Using the `hadoop fs` command, put the data into HDFS. This example places the file into an `/imported_data/` directory.

```
$ hadoop fs -put zones_frequency.tsv /imported_data/
```



3. Create and register a new HBase table in HCatalog, using the `hcat` command, passing it a DDL file to represent your table. You could also register an existing HBase table, using the same command. The DDL file format is specified as part of the [Hive REST API](#). The following example illustrates the basic mechanism.

```
CREATE TABLE
zones_frequency_table (id STRING, ngram STRING, year STRING, freq STRING, sources
STRING)
STORED BY 'org.apache.hcatalog.hbase.HBaseHCatStorageHandler'
TBLPROPERTIES (
  'hbase.table.name' = 'zones_frequency_table',
  'hbase.columns.mapping' = 'd:ngram,d:year,d:freq,d:sources',
  'hcat.hbase.output.bulkMode' = 'true'
);
```

```
$ hcat -f zones_frequency_table.ddl
```

4. Create a Pig file to process the TSV file created in step 1, using the DDL file created in step 3. Modify the file names and other parameters in this command to match your values if you use data different from this working example. `USING PigStorage('\t')` indicates that the input file is tab-delimited. For more details about Pig syntax, see the [Pig Latin](#) reference documentation.

```
A = LOAD 'hdfs:///imported_data/zones_frequency.tsv' USING PigStorage('\t') AS
(id:chararray, ngram:chararray, year:chararray, freq:chararray, sources:chararray);
-- DUMP A;
STORE A INTO 'zones_frequency_table' USING org.apache.hcatalog.pig.HCatStorer();
```

Save the file as `zones.bulkload.pig`.

5. Use the `pig` command to bulk-load the data into HBase.

```
$ pig -useHCatalog zones.bulkload.pig
```

The data is now in HBase and is available to use.

### Using the Java API

The Java API is the most common mechanism for getting data into HBase, through Put operations. The Thrift and REST APIs, as well as the HBase Shell, use the Java API. The following simple example uses the Java API to put data into an HBase table. The Java API traverses the entire write path and can cause compactions and region splits, which can adversely affect performance.

```
...
HTable table = null;
try {
  table = myCode.createTable(tableName, fam);
  int i = 1;
  List<Put> puts = new ArrayList<Put>();
  for (String labelExp : labelExps) {
    Put put = new Put(Bytes.toBytes("row" + i));
    put.add(fam, qual, HConstants.LATEST_TIMESTAMP, value);
    puts.add(put);
    i++;
  }
  table.put(puts);
} finally {
  if (table != null) {
    table.flushCommits();
  }
}
...
```

### Using the Apache Thrift Proxy API

The Apache Thrift library provides cross-language client-server remote procedure calls (RPCs), using *Thrift bindings*. A Thrift binding is client code generated by the Apache Thrift Compiler for a target language (such as

Python) that allows communication between the Thrift server and clients using that client code. HBase includes an Apache Thrift Proxy API, which allows you to write HBase applications in Python, C, C++, or another language that Thrift supports. The Thrift Proxy API is slower than the Java API and may have fewer features. To use the Thrift Proxy API, you need to configure and run the HBase Thrift server on your cluster. See [Installing and Starting the HBase Thrift Server](#). You also need to install the [Apache Thrift compiler](#) on your development system.

After the Thrift server is configured and running, generate Thrift bindings for the language of your choice, using an IDL file. A HBase IDL file named `HBase.thrift` is included as part of HBase. After generating the bindings, copy the Thrift libraries for your language into the same directory as the generated bindings. In the following Python example, these libraries provide the `thrift.transport` and `thrift.protocol` libraries. These commands show how you might generate the Thrift bindings for Python and copy the libraries on a Linux system.

```
$ mkdir HBaseThrift
$ cd HBaseThrift/
$ thrift -gen py /path/to/Hbase.thrift
$ mv gen-py/* .
$ rm -rf gen-py/
$ mkdir thrift
$ cp -rp ~/Downloads/thrift-0.9.0/lib/py/src/* ./thrift/
```

The following example shows a simple Python application using the Thrift Proxy API.

```
from thrift.transport import TSocket
from thrift.protocol import TBinaryProtocol
from thrift.transport import TTransport
from hbase import Hbase

# Connect to HBase Thrift server
transport = TTransport.TBufferedTransport(TSocket.TSocket(host, port))
protocol = TBinaryProtocol.TBinaryProtocolAccelerated(transport)

# Create and open the client connection
client = Hbase.Client(protocol)
transport.open()

# Modify a single row
mutations = [Hbase.Mutation(
    column='columnfamily:columndescriptor', value='columnvalue')]
client.mutateRow('tablename', 'rowkey', mutations)

# Modify a batch of rows
# Create a list of mutations per work of Shakespeare
mutationsbatch = []

for line in myDataFile:
    rowkey = username + "-" + filename + "-" + str(linenum).zfill(6)

    mutations = [
        Hbase.Mutation(column=messagecolumncf, value=line.strip()),
        Hbase.Mutation(column=linenumcolumncf, value=encode(linenum)),
        Hbase.Mutation(column=usernamecolumncf, value=username)
    ]

    mutationsbatch.append(Hbase.BatchMutation(row=rowkey, mutations=mutations))

# Run the mutations for all the lines in myDataFile
client.mutateRows(tablename, mutationsbatch)

transport.close()
```

The Thrift Proxy API does not support writing to HBase clusters that are secured using Kerberos.

This example was modified from the following two blog posts on <http://www.cloudera.com>. See them for more details.

- [Using the HBase Thrift Interface, Part 1](#)
- [Using the HBase Thrift Interface, Part 2](#)

## Using the REST Proxy API

After configuring and starting the [HBase REST Server](#) on your cluster, you can use the HBase REST Proxy API to stream data into HBase, from within another application or shell script, or by using an HTTP client such as `wget` or `curl`. The REST Proxy API is slower than the Java API and may have fewer features. This approach is simple and does not require advanced development experience to implement. However, like the Java and Thrift Proxy APIs, it uses the full write path and can cause compactions and region splits.

Specified addresses without existing data create new values. Specified addresses with existing data create new versions, overwriting an existing version if the row, column:qualifier, and timestamp all match that of the existing value.

```
$ curl -H "Content-Type: text/xml" http://localhost:8000/test/testrow/test:testcolumn
```

The REST Proxy API does not support writing to HBase clusters that are secured using Kerberos.

For full documentation and more examples, see the [REST Proxy API documentation](#).

## Using Flume

Apache Flume is a fault-tolerant system designed for ingesting data into HDFS, for use with Hadoop. You can configure Flume to write data directly into HBase. Flume includes two different *sinks* designed to work with HBase: `HBaseSink` (`org.apache.flume.sink.hbase.HBaseSink`) and `AsyncHBaseSink` (`org.apache.flume.sink.hbase.AsyncHBaseSink`). `HBaseSink` supports HBase IPC calls introduced in HBase 0.96, and allows you to write data to an HBase cluster that is secured by Kerberos, whereas `AsyncHBaseSink` does not. However, `AsyncHBaseSink` uses an asynchronous model and guarantees atomicity at the row level.

You configure `HBaseSink` and `AsyncHBaseSink` nearly identically. Following is an example configuration for each. Bold lines highlight differences in the configurations. For full documentation about configuring `HBaseSink` and `AsyncHBaseSink`, see the [Flume documentation](#). The `table`, `columnFamily`, and `column` parameters correlate to the HBase table, column family, and column where the data is to be imported. The serializer is the class that converts the data at the source into something HBase can use. Configure your sinks in the Flume configuration file.

In practice, you usually need to write your own serializer, which implements either `AsyncHBaseEventSerializer` or `HBaseEventSerializer`. The `HBaseEventSerializer` converts Flume Events into one or more HBase Puts, sends them to the HBase cluster, and is closed when the `HBaseSink` stops. `AsyncHBaseEventSerializer` starts and listens for Events. When it receives an Event, it calls the `setEvent` method and then calls the `getActions` and `getIncrements` methods. When the `AsyncHBaseSink` is stopped, the serializer `cleanUp` method is called. These methods return `PutRequest` and `AtomicIncrementRequest`, which are part of the `asynchbase` API.

### AsyncHBaseSink:

```
#Use the AsyncHBaseSink
host1.sinks.sink1.type = org.apache.flume.sink.hbase.AsyncHBaseSink
host1.sinks.sink1.channel = chl
host1.sinks.sink1.table = transactions
host1.sinks.sink1.columnFamily = clients
host1.sinks.sink1.column = charges
host1.sinks.sink1.batchSize = 5000
#Use the SimpleAsyncHbaseEventSerializer that comes with Flume
host1.sinks.sink1.serializer =
org.apache.flume.sink.hbase.SimpleAsyncHbaseEventSerializer
host1.sinks.sink1.serializer.incrementColumn = icol
host1.channels.chl.type=memory
```

### HBaseSink:

```
#Use the HBaseSink
host1.sinks.sink1.type = org.apache.flume.sink.hbase.HBaseSink
host1.sinks.sink1.channel = chl
host1.sinks.sink1.table = transactions
host1.sinks.sink1.columnFamily = clients
host1.sinks.sink1.column = charges
```

```
host1.sinks.sink1.batchSize = 5000
#Use the SimpleHbaseEventSerializer that comes with Flume
host1.sinks.sink1.serializer = org.apache.flume.sink.hbase.SimpleHbaseEventSerializer
host1.sinks.sink1.serializer.incrementColumn = icol
host1.channels.ch1.type=memory
```

The following serializer, taken from an [Apache Flume blog post by Dan Sandler](#), splits the event body based on a delimiter and inserts each split into a different column. The row is defined in the event header. When each event is received, a counter is incremented to track the number of events received.

```
/**
 * A serializer for the AsyncHBaseSink, which splits the event body into
 * multiple columns and inserts them into a row whose key is available in
 * the headers
 */
public class SplittingSerializer implements AsyncHbaseEventSerializer {
    private byte[] table;
    private byte[] colFam;
    private Event currentEvent;
    private byte[][] columnNames;
    private final List<PutRequest> puts = new ArrayList<PutRequest>();
    private final List<AtomicIncrementRequest> incs = new
ArrayList<AtomicIncrementRequest>();
    private byte[] currentRowKey;
    private final byte[] eventCountCol = "eventCount".getBytes();    @Override
    public void initialize(byte[] table, byte[] cf) {
        this.table = table;
        this.colFam = cf;
    }
    @Override
    public void setEvent(Event event) {
        // Set the event and verify that the rowKey is not present
        this.currentEvent = event;
        String rowKeyStr = currentEvent.getHeaders().get("rowKey");
        if (rowKeyStr == null) {
            throw new FlumeException("No row key found in headers!");
        }
        currentRowKey = rowKeyStr.getBytes();
    }
    @Override
    public List<PutRequest> getActions() {
        // Split the event body and get the values for the columns
        String eventStr = new String(currentEvent.getBody());
        String[] cols = eventStr.split(",");
        puts.clear();
        for (int i = 0; i < cols.length; i++) {
            //Generate a PutRequest for each column.
            PutRequest req = new PutRequest(table, currentRowKey, colFam,
                columnNames[i], cols[i].getBytes());
            puts.add(req);
        }
        return puts;
    }
    @Override
    public List<AtomicIncrementRequest> getIncrements() {
        incs.clear();
        //Increment the number of events received
        incs.add(new AtomicIncrementRequest(table, "totalEvents".getBytes(), colFam,
eventCountCol));
        return incs;
    }
    @Override
    public void cleanUp() {
        table = null;
        colFam = null;
        currentEvent = null;
        columnNames = null;
        currentRowKey = null;
    }
    @Override
    public void configure(Context context) {
        //Get the column names from the configuration
        String cols = new String(context.getString("columns"));
        String[] names = cols.split(",");
        byte[][] columnNames = new byte[names.length][];
        int i = 0;
        for(String name : names) {
```

```

        columnNames[i++] = name.getBytes();
    }
    @Override
    public void configure(ComponentConfiguration conf) {
    }
}

```

### Using Spark

You can write data to HBase from Apache Spark by using `*def saveAsHadoopDataset(conf: JobConf): Unit*`. This example is adapted from [a post on the spark-users mailing list](#).

```

// Note: mapred package is used, instead of the mapreduce package which
contains new hadoop APIs.

*import org.apache.hadoop.hbase.mapred.TableOutputFormat*
*import org.apache.hadoop.hbase.client._*
// ... some other settings

*val conf = HBaseConfiguration.create()*

// general hbase setting
*conf.set("hbase.rootdir", "hdfs://" + nameNodeURL + ":" + hdfsPort +
"/hbase")*
*conf.setBoolean("hbase.cluster.distributed", true)*
*conf.set("hbase.zookeeper.quorum", hostname)*
*conf.setInt("hbase.client.scanner.caching", 10000)*
// ... some other settings

*val jobConfig: JobConf = new JobConf(conf, this.getClass)*

// Note: TableOutputFormat is used as deprecated code, because JobConf is
an old hadoop API
*jobConfig.setOutputFormat(classOf[TableOutputFormat])*
*jobConfig.set(TableOutputFormat.OUTPUT_TABLE, outputTable)*

```

Next, provide the mapping between how the data looks in Spark and how it should look in HBase. The following example assumes that your HBase table has two column families, `col_1` and `col_2`, and that your data is formatted in sets of three in Spark, like `(row_key, col_1, col_2)`.

```

*def convert(triple: (Int, Int, Int)) = {*
*   val p = new Put(Bytes.toBytes(triple._1))*
*   p.add(Bytes.toBytes("cf"), Bytes.toBytes("col_1"),
Bytes.toBytes(triple._2))*
*   p.add(Bytes.toBytes("cf"), Bytes.toBytes("col_2"),
Bytes.toBytes(triple._3))*
*   (new ImmutableBytesWritable, p)*
*}*

```

To write the data from Spark to HBase, you might use:

```

*new
PairRDDFunctions(localData.map(convert)).saveAsHadoopDataset(jobConfig)*

```

### Using Spark and Kafka

This example, written in Scala, uses Apache Spark in conjunction with the Apache Kafka message bus to stream data from Spark to HBase. The example was provided in [SPARK-944](#). It produces some random words and then stores them in an HBase table, creating the table if necessary.

```

package org.apache.spark.streaming.examples

import java.util.Properties

import kafka.producer._

import org.apache.hadoop.hbase.{ HBaseConfiguration, HColumnDescriptor, HTableDescriptor
}

```

```

import org.apache.hadoop.hbase.client.{ HBaseAdmin, Put }
import org.apache.hadoop.hbase.io.ImmutableBytesWritable
import org.apache.hadoop.hbase.mapred.TableOutputFormat
import org.apache.hadoop.hbase.mapreduce.TableInputFormat
import org.apache.hadoop.hbase.util.Bytes
import org.apache.hadoop.mapred.JobConf
import org.apache.spark.SparkContext
import org.apache.spark.rdd.{ PairRDDFunctions, RDD }
import org.apache.spark.streaming._
import org.apache.spark.streaming.StreamingContext._
import org.apache.spark.streaming.kafka._

object MetricAggregatorHBase {
  def main(args : Array[String]) {
    if (args.length < 6) {
      System.err.println("Usage: MetricAggregatorTest <master> <zkQuorum> <group>
<topics> <destHBaseTableName> <numThreads>")
      System.exit(1)
    }

    val Array(master, zkQuorum, group, topics, hbaseTableName, numThreads) = args

    val conf = HBaseConfiguration.create()
    conf.set("hbase.zookeeper.quorum", zkQuorum)

    // Initialize hBase table if necessary
    val admin = new HBaseAdmin(conf)
    if (!admin.isTableAvailable(hbaseTableName)) {
      val tableDesc = new HTableDescriptor(hbaseTableName)
      tableDesc.addFamily(new HColumnDescriptor("metric"))
      admin.createTable(tableDesc)
    }

    // setup streaming context
    val ssc = new StreamingContext(master, "MetricAggregatorTest", Seconds(2),
      System.getenv("SPARK_HOME"), StreamingContext.jarOfClass(this.getClass))
    ssc.checkpoint("checkpoint")

    val topiccpMap = topics.split(",").map((_, numThreads.toInt)).toMap
    val lines = KafkaUtils.createStream(ssc, zkQuorum, group, topiccpMap)
      .map { case (key, value) => ((key, Math.floor(System.currentTimeMillis() /
60000).toLong * 60), value.toInt) }

    val aggr = lines.reduceByKeyAndWindow(add _, Minutes(1), Minutes(1), 2)

    aggr.foreach(line => saveToHBase(line, zkQuorum, hbaseTableName))

    ssc.start

    ssc.awaitTermination
  }

  def add(a : Int, b : Int) = { (a + b) }

  def saveToHBase(rdd : RDD[(String, Long), Int], zkQuorum : String, tableName :
String) = {
    val conf = HBaseConfiguration.create()
    conf.set("hbase.zookeeper.quorum", zkQuorum)

    val jobConfig = new JobConf(conf)
    jobConfig.set(TableOutputFormat.OUTPUT_TABLE, tableName)
    jobConfig.setOutputFormat(classOf[TableOutputFormat])

    new PairRDDFunctions(rdd.map { case ((metricId, timestamp), value) =>
createHBaseRow(metricId, timestamp, value) }).saveAsHadoopDataset(jobConfig)
  }

  def createHBaseRow(metricId : String, timestamp : Long, value : Int) = {
    val record = new Put(Bytes.toBytes(metricId + "~" + timestamp))

    record.add(Bytes.toBytes("metric"), Bytes.toBytes("col"),
Bytes.toBytes(value.toString))
  }

```

```

    (new ImmutableBytesWritable, record)
  }
}

// Produces some random words between 1 and 100.
object MetricDataProducer {

  def main(args : Array[String]) {
    if (args.length < 2) {
      System.err.println("Usage: MetricDataProducer <metadataBrokerList> <topic>
<messagesPerSec>")
      System.exit(1)
    }

    val Array(brokers, topic, messagesPerSec) = args

    // ZooKeeper connection properties
    val props = new Properties()
    props.put("metadata.broker.list", brokers)
    props.put("serializer.class", "kafka.serializer.StringEncoder")

    val config = new ProducerConfig(props)
    val producer = new Producer[String, String](config)

    // Send some messages
    while (true) {
      val messages = (1 to messagesPerSec.toInt).map { messageNum =>
        {
          val metricId = scala.util.Random.nextInt(10)
          val value = scala.util.Random.nextInt(1000)
          new KeyedMessage[String, String](topic, metricId.toString, value.toString)
        }
      }.toArray

      producer.send(messages : _*)
      Thread.sleep(100)
    }
  }
}

```

### Using a Custom MapReduce Job

Many of the methods to import data into HBase use MapReduce implicitly. If none of those approaches fit your needs, you can use MapReduce directly to convert data to a series of HFiles or API calls for import into HBase. In this way, you can import data from Avro, Parquet, or another format into HBase, or export data from HBase into another format, using API calls such as [TableOutputFormat](#), [HFileOutputFormat](#), and [TableInputFormat](#).

### Checking and Repairing HBase Tables

HBaseFsck (hbck) is a command-line tool that checks for region consistency and table integrity problems and repairs corruption. It works in two basic modes — a read-only inconsistency identifying mode and a multi-phase read-write repair mode.

- **Read-only inconsistency identification:** In this mode, which is the default, a report is generated but no repairs are attempted.
- **Read-write repair mode:** In this mode, if errors are found, hbck attempts to repair them.

You can run hbck manually or configure the hbck poller to run hbck periodically.

#### Running hbck Manually

The hbck command is located in the bin directory of the HBase install.

- With no arguments, hbck checks HBase for inconsistencies and prints OK if no inconsistencies are found, or the number of inconsistencies otherwise.
- With the `-details` argument, hbck checks HBase for inconsistencies and prints a detailed report.
- To limit hbck to only checking specific tables, provide them as a space-separated list: `hbck <table1> <table2>`



- If region-level inconsistencies are found, use the `-fix` argument to direct `hbck` to try to fix them. The following sequence of steps is followed:
  1. The standard check for inconsistencies is run.
  2. If needed, repairs are made to tables.
  3. If needed, repairs are made to regions. Regions are closed during repair.
- You can also fix individual region-level inconsistencies separately, rather than fixing them automatically with the `-fix` argument.
  - `-fixAssignments` repairs unassigned, incorrectly assigned or multiply assigned regions.
  - `-fixMeta` removes rows from `hbase:meta` when their corresponding regions are not present in HDFS and adds new meta rows if regions are present in HDFS but not in `hbase:meta`.
  - `-repairHoles` creates HFiles for new empty regions on the filesystem and ensures that the new regions are consistent.
  - `-fixHdfsOrphans` repairs a region directory that is missing a region metadata file (the `.regioninfo` file).
  - `-fixHdfsOverlaps` fixes overlapping regions. You can further tune this argument using the following options:
    - `-maxMerge <n>` controls the maximum number of regions to merge.
    - `-sidelineBigOverlaps` attempts to sideline the regions which overlap the largest number of other regions.
    - `-maxOverlapsToSideline <n>` limits the maximum number of regions to sideline.
- To try to repair all inconsistencies and corruption at once, use the `-repair` option, which includes all the region and table consistency options.

For more details about the `hbck` command, see [Appendix C](#) of the HBase Reference Guide.

### Configuring the `hbck` Poller

The `hbck` poller is a feature of Cloudera Manager, which can be configured to run `hbck` automatically, in read-only mode, and send alerts if errors are found. By default, it runs every 30 minutes. Several configuration settings are available for the `hbck` poller. The `hbck` poller is not provided if you use CDH without Cloudera Manager.

### Configuring the `hbck` Poller

1. Go to the HBase service and click **Configuration**.
2. Configure the alert behavior with the following settings:
  - **HBase Hbck Poller Maximum Error Count:** The maximum number of errors that the `hbck` poller will retain through a given run.
  - **HBase Hbck Region Error Count:** An alert is published if at least this number of regions is detected with errors across all regions in this service. If the value is not set, alerts will not be published based on the count of regions with errors.
  - **Alert Threshold:** An alert is published if the number of errors reaches this threshold.
  - **HBase Hbck Error Count Alert Threshold:** An alert is published if at least this number of tables is detected with errors across all tables in this service. Some errors are not associated with a region, such as `RS_CONNECT_FAILURE`. If the value is not set, alerts will not be published based on the count of tables with errors.
  - **HBase Hbck Alert Error Codes:** An alert is published errors match any of the specified codes. The default behavior is not to limit the error codes which trigger an alert. May be set to one or more of the following:
    - UNKNOWN
    - NO\_META\_REGION
    - NULL\_ROOT\_REGION
    - NO\_VERSION\_FILE



- NOT\_IN\_META\_HDFS
- NOT\_IN\_META
- NOT\_IN\_META\_OR\_DEPLOYED
- NOT\_IN\_HDFS\_OR\_DEPLOYED
- NOT\_IN\_HDFS
- SERVER\_DOES\_NOT\_MATCH\_META
- NOT\_DEPLOYED
- MULTI\_DEPLOYED
- SHOULD\_NOT\_BE\_DEPLOYED
- MULTI\_META\_REGION
- RS\_CONNECT\_FAILURE
- FIRST\_REGION\_STARTKEY\_NOT\_EMPTY
- LAST\_REGION\_ENDKEY\_NOT\_EMPTY
- DUPE\_STARTKEYS
- HOLE\_IN\_REGION\_CHAIN
- OVERLAP\_IN\_REGION\_CHAIN
- REGION\_CYCLE
- DEGENERATE\_REGION
- ORPHAN\_HDFS\_REGION
- LINGERING\_SPLIT\_PARENT
- NO\_TABLEINFO\_FILE

3. To configure the polling interval, edit the **Service Monitor Derived Configs Advanced Configuration Snippet** with a setting such as the following, which sets the polling interval to 60 minutes. Restart the RegionServers for the changes to take effect.

```
<property>
  <name>smon.hbase.fsckpoller.interval.ms</name>
  <value>3600000</value>
</property>
```

## Configuring Hive

Use the following procedures to configure HiveServer2 and the Hive MetaStore. To configure high availability for the Hive MetaStore, see [Hive Metastore High Availability](#) on page 258.

### Configuring Heap Size and Garbage Collection for Hive Components

HiveServer2 and the Hive metastore require sufficient memory in order to run correctly. The default heap size of 256 MB for each component is inadequate for production workloads. Consider the following guidelines for sizing the heap for each component, based upon your cluster size.

**Table 4: Hive Heap Size Recommendations**

Cluster Size	HiveServer2 Heap Size	Hive Metastore Heap Size
100 nodes or larger	24 GB	24 GB
50-99 nodes	12 GB	12 GB
11-49 nodes	6 GB	6 GB
2-10 nodes	2 GB	2 GB
1 node	256 MB	256 MB

## Configuring CDH and Managed Services

In addition, workstations running The Beehive CLI should use a heap size of at least 2 GB.

### Configuring Heap Size and Garbage Collection for Hive Components

To configure the heap size for HiveServer2 and Hive metastore, use the `hive-env.sh` advanced configuration snippet if you use Cloudera Manager, or edit `/etc/hive/hive-env.sh` otherwise, and set the `-Xmx` parameter in the `HADOOP_OPTS` variable to the desired maximum heap size.

To configure the heap size for the Beehive CLI, use the `hive-env.sh` advanced configuration snippet if you use Cloudera Manager, or edit `/etc/hive/hive-env.sh` otherwise, and set the `HADOOP_HEAPSIZE` environment variable before starting the Beehive CLI.

The following example shows a configuration with the following settings:

- HiveServer2 uses 12 GB heap
- Hive metastore heap uses 12 GB heap
- Hive clients use 2 GB heap

The settings to change are in bold. All of these lines are commented out (prefixed with a `#` character) by default. Uncomment the lines by removing the `#` character.

```
if [ "$SERVICE" = "cli" ]; then
  if [ -z "$DEBUG" ]; then
    export HADOOP_OPTS="$HADOOP_OPTS -XX:NewRatio=12 -Xmx12288m -Xms10m
-XX:MaxHeapFreeRatio=40 -XX:MinHeapFreeRatio=15 -XX:+useParNewGC -XX:-useGCOverheadLimit"

  else
    export HADOOP_OPTS="$HADOOP_OPTS -XX:NewRatio=12 -Xmx12288m -Xms10m
-XX:MaxHeapFreeRatio=40 -XX:MinHeapFreeRatio=15 -XX:-useGCOverheadLimit"
  fi
fi
export HADOOP_HEAPSIZE=2048
```

You can choose whether to use the Concurrent Collector or the New Parallel Collector for garbage collection, by passing `-XX:+useParNewGC` or `-XX:+useConcMarkSweepGC` in the `HADOOP_OPTS` lines above, and you can tune the garbage collection overhead limit by setting `-XX:-useGCOverheadLimit`. To enable the garbage collection overhead limit, remove the setting or change it to `-XX:+useGCOverheadLimit`.

### Configuration for WebHCat

If you want to use WebHCat, you need to set the `PYTHON_CMD` variable in `/etc/default/hive-webhcat-server` after installing Hive; for example:

```
export PYTHON_CMD=/usr/bin/python
```

## Configuring Impala

This section explains how to configure Impala to accept connections from applications that use popular programming APIs:

- [Post-Installation Configuration for Impala](#) on page 163
- [Configuring Impala to Work with ODBC](#) on page 166
- [Configuring Impala to Work with JDBC](#) on page 168

This type of configuration is especially useful when using Impala in combination with Business Intelligence tools, which use these standard interfaces to query different kinds of database and Big Data systems.

You can also configure these other aspects of Impala:

- [Overview of Impala Security](#)
- [Modifying Impala Startup Options](#)

## Post-Installation Configuration for Impala

This section describes the mandatory and recommended configuration settings for Cloudera Impala. If Impala is installed using Cloudera Manager, some of these configurations are completed automatically; you must still configure short-circuit reads manually. If you installed Impala without Cloudera Manager, or if you want to customize your environment, consider making the changes described in this topic.

In some cases, depending on the level of Impala, CDH, and Cloudera Manager, you might need to add particular component configuration details in one of the free-form fields on the Impala configuration pages within Cloudera Manager. In Cloudera Manager 4, these fields are labelled **Safety Valve**; in Cloudera Manager 5, they are called **Advanced Configuration Snippet**.

- You must enable short-circuit reads, whether or not Impala was installed through Cloudera Manager. This setting goes in the Impala configuration settings, not the Hadoop-wide settings.
- If you installed Impala in an environment that is not managed by Cloudera Manager, you must enable block location tracking, and you can optionally enable native checksumming for optimal performance.
- If you deployed Impala using Cloudera Manager see [Testing Impala Performance](#) to confirm proper configuration.

### Mandatory: Short-Circuit Reads

Enabling short-circuit reads allows Impala to read local data directly from the file system. This removes the need to communicate through the DataNodes, improving performance. This setting also minimizes the number of additional copies of data. Short-circuit reads requires `libhadoop.so` (the Hadoop Native Library) to be accessible to both the server and the client. `libhadoop.so` is not available if you have installed from a tarball. You must install from an `.rpm`, `.deb`, or `parcel` to use short-circuit local reads.

- **Note:** If you use Cloudera Manager, you can enable short-circuit reads through a checkbox in the user interface and that setting takes effect for Impala as well.

Cloudera strongly recommends using Impala with CDH 4.2 or higher, ideally the latest 4.x release. Impala does support short-circuit reads with CDH 4.1, but for best performance, upgrade to CDH 4.3 or higher. The process of configuring short-circuit reads varies according to which version of CDH you are using. Choose the procedure that is appropriate for your environment.

#### To configure DataNodes for short-circuit reads with CDH 4.2 or higher:

1. Copy the client `core-site.xml` and `hdfs-site.xml` configuration files from the Hadoop configuration directory to the Impala configuration directory. The default Impala configuration location is `/etc/impala/conf`.
2. On all Impala nodes, configure the following properties in Impala's copy of `hdfs-site.xml` as shown:

```
<property>
  <name>dfs.client.read.shortcircuit</name>
  <value>true</value>
</property>

<property>
  <name>dfs.domain.socket.path</name>
  <value>/var/run/hdfs-sockets/dn</value>
</property>

<property>
  <name>dfs.client.file-block-storage-locations.timeout.millis</name>
  <value>10000</value>
</property>
```

3. If `/var/run/hadoop-hdfs/` is group-writable, make sure its group is `root`.

- **Note:** If you are also going to enable block location tracking, you can skip copying configuration files and restarting DataNodes and go straight to [Optional: Block Location Tracking](#). Configuring short-circuit reads and block location tracking require the same process of copying files and restarting services, so you can complete that process once when you have completed all configuration changes. Whether you copy files and restart services now or during configuring block location tracking, short-circuit reads are not enabled until you complete those final steps.

4. After applying these changes, restart all DataNodes.

**To configure DataNodes for short-circuit reads with CDH 4.1:**

- **Note:** Cloudera strongly recommends using Impala with CDH 4.2 or higher, ideally the latest 4.x release. Impala does support short-circuit reads with CDH 4.1, but for best performance, upgrade to CDH 4.3 or higher. The process of configuring short-circuit reads varies according to which version of CDH you are using. Choose the procedure that is appropriate for your environment.

1. Enable short-circuit reads by adding settings to the Impala `core-site.xml` file.

- If you installed Impala using Cloudera Manager, short-circuit reads should be properly configured, but you can review the configuration by checking the contents of the `core-site.xml` file, which is installed at `/etc/impala/conf` by default.
- If you installed using packages, instead of using Cloudera Manager, create the `core-site.xml` file. This can be easily done by copying the `core-site.xml` client configuration file from another machine that is running Hadoop services. This file must be copied to the Impala configuration directory. The Impala configuration directory is set by the `IMPALA_CONF_DIR` environment variable and is by default `/etc/impala/conf`. To confirm the Impala configuration directory, check the `IMPALA_CONF_DIR` environment variable value.

- **Note:** If the Impala configuration directory does not exist, create it and then add the `core-site.xml` file.

Add the following to the `core-site.xml` file:

```
<property>
  <name>dfs.client.read.shortcircuit</name>
  <value>true</value>
</property>
```

- **Note:** For an installation managed by Cloudera Manager, specify these settings in the Impala dialogs, in the options field for HDFS. In Cloudera Manager 4, these fields are labelled **Safety Valve**; in Cloudera Manager 5, they are called **Advanced Configuration Snippet**.

2. For each DataNode, enable access by adding the following to the `hdfs-site.xml` file:

```
<property>
  <name>dfs.client.use.legacy.blockreader.local</name>
  <value>true</value>
</property>

<property>
  <name>dfs.datanode.data.dir.perm</name>
  <value>750</value>
</property>

<property>
  <name>dfs.block.local-path-access.user</name>
  <value>impala</value>
</property>
```

```
<property>
  <name>dfs.client.file-block-storage-locations.timeout.millis</name>
  <value>10000</value>
</property>
```

- **Note:** In the preceding example, the `dfs.block.local-path-access.user` is the user running the `impalad` process. By default, that account is `impala`.

3. Use `usermod` to add users requiring local block access to the appropriate HDFS group. For example, if you assigned `impala` to the `dfs.block.local-path-access.user` property, you would add `impala` to the `hadoop` HDFS group:

```
$ usermod -a -G hadoop impala
```

- **Note:** The default HDFS group is `hadoop`, but it is possible to have an environment configured to use an alternate group. To find the configured HDFS group name using the Cloudera Manager admin console, click **Services** and click **HDFS**. Click the **Configuration** tab. Under **Service-Wide**, click **Advanced** in the left column. The **Shared Hadoop Group Name** property contains the group name.

- **Note:** If you are going to enable block location tracking, you can skip copying configuration files and restarting DataNodes and go straight to [Mandatory: Block Location Tracking](#) on page 165. Configuring short-circuit reads and block location tracking require the same process of copying files and restarting services, so you can complete that process once when you have completed all configuration changes. Whether you copy files and restart services now or during configuring block location tracking, short-circuit reads are not enabled until you complete those final steps.

4. Copy the client `core-site.xml` and `hdfs-site.xml` configuration files from the Hadoop configuration directory to the Impala configuration directory. The default Impala configuration location is `/etc/impala/conf`.
5. After applying these changes, restart all DataNodes.

### Mandatory: Block Location Tracking

Enabling block location metadata allows Impala to know which disk data blocks are located on, allowing better utilization of the underlying disks. Impala will not start unless this setting is enabled.

#### To enable block location tracking:

1. For each DataNode, adding the following to the `hdfs-site.xml` file:

```
<property>
  <name>dfs.datanode.hdfs-blocks-metadata.enabled</name>
  <value>true</value>
</property>
```

2. Copy the client `core-site.xml` and `hdfs-site.xml` configuration files from the Hadoop configuration directory to the Impala configuration directory. The default Impala configuration location is `/etc/impala/conf`.
3. After applying these changes, restart all DataNodes.

### Optional: Native Checksumming

Enabling native checksumming causes Impala to use an optimized native library for computing checksums, if that library is available.

#### To enable native checksumming:

If you installed CDH from packages, the native checksumming library is installed and setup correctly. In such a case, no additional steps are required. Conversely, if you installed by other means, such as with tarballs, native

checksumming may not be available due to missing shared objects. Finding the message "Unable to load native-hadoop library for your platform... using builtin-java classes where applicable" in the Impala logs indicates native checksumming may be unavailable. To enable native checksumming, you must build and install `libhadoop.so` (the [Hadoop Native Library](#)).

### Configuring Impala to Work with ODBC

Third-party products can be designed to integrate with Impala using ODBC. For the best experience, ensure any third-party product you intend to use is supported. Verifying support includes checking that the versions of Impala, ODBC, the operating system, and the third-party product have all been approved for use together. Before configuring your systems to use ODBC, download a connector.

- **Note:** You may need to sign in and accept license agreements before accessing the pages required for downloading ODBC connectors.

Versions 2.5 and 2.0 of the Cloudera ODBC Connector, currently certified for some but not all BI applications, use the HiveServer2 protocol, corresponding to Impala port 21050. Impala supports Kerberos authentication with all the supported versions of the driver, and requires ODBC 2.05.13 for Impala or higher for LDAP username/password authentication.

Version 1.x of the Cloudera ODBC Connector uses the original HiveServer1 protocol, corresponding to Impala port 21000.

See the [downloads page](#) for the versions of these drivers for different products, and the [documentation page](#) for installation instructions.

- **Important:** If you are using the Cloudera Connector for Tableau, to connect Impala to your Kerberos-secured CDH clusters, contact your Tableau account representative for an updated Tableau Data-connection Customization (TDC) file. The updated TDC file will override the Tableau connection settings to set specific parameters on the connection string that are required for a secure connection.

- **Note:** If your JDBC or ODBC application connects to Impala through a load balancer such as `haproxy`, be cautious about reusing the connections. If the load balancer has set up connection timeout values, either check the connection frequently so that it never sits idle longer than the load balancer timeout value, or check the connection validity before using it and create a new one if the connection has been closed.

To illustrate the outline of the setup process, here is a transcript of a session to set up all required drivers and a business intelligence application that uses the ODBC driver, under Mac OS X. Each `.dmg` file runs a GUI-based installer, first for the [underlying JDBC driver](#) needed for non-Windows systems, then for the Cloudera ODBC Connector, and finally for the BI tool itself.

```
$ ls -l
Cloudera-ODBC-Driver-for-Impala-Install-Guide.pdf
BI_Tool_Installer.dmg
iodbc-sdk-3.52.7-macosx-10.5.dmg
ClouderaImpalaODBC.dmg
$ open iodbc-sdk-3.52.7-macosx-10.5.dmg
Install the IODBC driver using its installer
$ open ClouderaImpalaODBC.dmg
Install the Cloudera ODBC Connector using its installer
$ installer_dir=$(pwd)
$ cd /opt/cloudera/impalaodbc
$ ls -l
Cloudera ODBC Driver for Impala Install Guide.pdf
Readme.txt
Setup
lib
ErrorMessages
Release Notes.txt
Tools
```

```

$ cd Setup
$ ls
odbc.ini odbcinst.ini
$ cp odbc.ini ~/.odbc.ini
$ vi ~/.odbc.ini
$ cat ~/.odbc.ini
[ODBC]
# Specify any global ODBC configuration here such as ODBC tracing.

[ODBC Data Sources]
Sample Cloudera Impala DSN=Cloudera ODBC Driver for Impala

[Sample Cloudera Impala DSN]

# Description: DSN Description.
# This key is not necessary and is only to give a description of the data source.
Description=Cloudera ODBC Driver for Impala DSN

# Driver: The location where the ODBC driver is installed to.
Driver=/opt/cloudera/impalaodbc/lib/universal/libclouderaimpalaodbc.dylib

# The DriverUnicodeEncoding setting is only used for SimbaDM
# When set to 1, SimbaDM runs in UTF-16 mode.
# When set to 2, SimbaDM runs in UTF-8 mode.
#DriverUnicodeEncoding=2

# Values for HOST, PORT, KrbFQDN, and KrbServiceName should be set here.
# They can also be specified on the connection string.
HOST=hostname.sample.example.com
PORT=21050
Schema=default

# The authentication mechanism.
# 0 - No authentication (NOSASL)
# 1 - Kerberos authentication (SASL)
# 2 - Username authentication (SASL)
# 3 - Username/password authentication (SASL)
# 4 - Username/password authentication with SSL (SASL)
# 5 - No authentication with SSL (NOSASL)
# 6 - Username/password authentication (NOSASL)
AuthMech=0

# Kerberos related settings.
KrbFQDN=
KrbRealm=
KrbServiceName=

# Username/password authentication with SSL settings.
UID=
PWD=
CAIssuedCertNamesMismatch=1
TrustedCerts=/opt/cloudera/impalaodbc/lib/universal/cacerts.pem

# Specify the proxy user ID to use.
#DelegationUID=

# General settings
TSaslTransportBufSize=1000
RowsFetchedPerBlock=10000
SocketTimeout=0
StringColumnLength=32767
UseNativeQuery=0
$ pwd
/opt/cloudera/impalaodbc/Setup
$ cd $installer_dir
$ open BI_Tool_Installer.dmg
Install the BI tool using its installer
$ ls /Applications | grep BI_Tool
BI_Tool.app
$ open -a BI_Tool.app
In the BI tool, connect to a data source using port 21050

```

### Configuring Impala to Work with JDBC

Impala supports the standard JDBC interface, allowing access from commercial Business Intelligence tools and custom software written in Java or other programming languages. The JDBC driver allows you to access Impala from a Java program that you write, or a Business Intelligence or similar tool that uses JDBC to communicate with various database products.

Setting up a JDBC connection to Impala involves the following steps:

- Verifying the communication port where the Impala daemons in your cluster are listening for incoming JDBC requests.
- Installing the JDBC driver on every system that runs the JDBC-enabled application.
- Specifying a connection string for the JDBC application to access one of the servers running the `impalad` daemon, with the appropriate security settings.

#### Configuring the JDBC Port

The default port used by JDBC 2.0 and later (as well as ODBC 2.x) is 21050. Impala server accepts JDBC connections through this same port 21050 by default. Make sure this port is available for communication with other hosts on your network, for example, that it is not blocked by firewall software. If your JDBC client software connects to a different port, specify that alternative port number with the `--hs2_port` option when starting `impalad`. See [Starting Impala](#) for details about Impala startup options. See [Ports Used by Impala](#) for information about all ports used for communication between Impala and clients or between Impala components.

#### Choosing the JDBC Driver

In Impala 2.0 and later, you have the choice between the Cloudera JDBC Connector and the Hive 0.13 JDBC driver. Cloudera recommends using the Cloudera JDBC Connector where practical.

If you are already using JDBC applications with an earlier Impala release, you must update your JDBC driver to one of these choices, because the Hive 0.12 driver that was formerly the only choice is not compatible with Impala 2.0 and later.

You download and install the Cloudera JDBC 2.5 connector on any Linux, Windows, or Mac system where you intend to run JDBC-enabled applications. From the [Cloudera Connectors download page](#), you choose the appropriate protocol (JDBC or ODBC) and target product (Impala or Hive). The ease of downloading and installing on non-CDH systems makes this connector a convenient choice for organizations with heterogeneous environments.

You install the Hive JDBC driver (`hive-jdbc` package) through the Linux package manager, on hosts within the CDH cluster.

Both the Hive JDBC driver and the Cloudera JDBC 2.5 Connector provide a substantial speed increase for JDBC applications with Impala 2.0 and higher, for queries that return large result sets.

#### Enabling Impala JDBC Support on Client Systems

The Impala JDBC integration is made possible by a client-side JDBC driver, made up of several Java JAR files. The same driver is used by Impala and Hive.

To get the JAR files, install the Hive JDBC driver on each CDH-enabled host in the cluster that will run JDBC applications. Follow the instructions for [CDH 5](#) or [CDH 4](#).

- **Note:** The latest JDBC driver, corresponding to Hive 0.13, provides substantial performance improvements for Impala queries that return large result sets. Impala 2.0 and later are compatible with the Hive 0.13 driver. If you already have an older JDBC driver installed, and are running Impala 2.0 or higher, consider upgrading to the latest Hive JDBC driver for best performance with JDBC applications.



If you are using JDBC-enabled applications on hosts outside the CDH cluster, you cannot use the CDH install procedure on the non-CDH hosts. Install the JDBC driver on at least one CDH host using the preceding procedure. Then download the JAR files to each client machine that will use JDBC with Impala:

```
commons-logging-X.X.X.jar
hadoop-common.jar
hive-common-X.XX.X-cdhX.X.X.jar
hive-jdbc-X.XX.X-cdhX.X.X.jar
hive-metastore-X.XX.X-cdhX.X.X.jar
hive-service-X.XX.X-cdhX.X.X.jar
httpclient-X.X.X.jar
httpcore-X.X.X.jar
libfb303-X.X.X.jar
libthrift-X.X.X.jar
log4j-X.X.XX.jar
slf4j-api-X.X.X.jar
slf4j-log4j-X.X.X.jar
```

To enable JDBC support for Impala on the system where you run the JDBC application:

1. Download the JAR files listed above to each client machine.

- **Note:** For Maven users, see [this sample github page](#) for an example of the dependencies you could add to a `pom` file instead of downloading the individual JARs.

2. Store the JAR files in a location of your choosing, ideally a directory already referenced in your `CLASSPATH` setting. For example:
  - On Linux, you might use a location such as `/opt/jars/`.
  - On Windows, you might use a subdirectory underneath `C:\Program Files`.
3. To successfully load the Impala JDBC driver, client programs must be able to locate the associated JAR files. This often means setting the `CLASSPATH` for the client process to include the JARs. Consult the documentation for your JDBC client for more details on how to install new JDBC drivers, but some examples of how to set `CLASSPATH` variables include:

- On Linux, if you extracted the JARs to `/opt/jars/`, you might issue the following command to prepend the JAR files path to an existing classpath:

```
export CLASSPATH=/opt/jars/*.jar:$CLASSPATH
```

- On Windows, use the **System Properties** control panel item to modify the **Environment Variables** for your system. Modify the environment variables to include the path to which you extracted the files.

- **Note:** If the existing `CLASSPATH` on your client machine refers to some older version of the Hive JARs, ensure that the new JARs are the first ones listed. Either put the new JAR files earlier in the listings, or delete the other references to Hive JAR files.

## Establishing JDBC Connections

The JDBC driver class is `org.apache.hive.jdbc.HiveDriver`. Once you have configured Impala to work with JDBC, you can establish connections between the two. To do so for a cluster that does not use Kerberos authentication, use a connection string of the form `jdbc:hive2://host:port/;auth=noSasl`. For example, you might use:

```
jdbc:hive2://myhost.example.com:21050/;auth=noSasl
```

## Configuring CDH and Managed Services

To connect to an instance of Impala that requires Kerberos authentication, use a connection string of the form `jdbc:hive2://host:port/;principal=principal_name`. The principal must be the same user principal you used when starting Impala. For example, you might use:

```
jdbc:hive2://myhost.example.com:21050/;principal=impala/myhost.example.com@H2.EXAMPLE.COM
```

To connect to an instance of Impala that requires LDAP authentication, use a connection string of the form `jdbc:hive2://host:port/db_name;user=ldap_userid;password=ldap_password`. For example, you might use:

```
jdbc:hive2://myhost.example.com:21050/test_db;user=fred;password=xyz123
```

- **Note:** If your JDBC or ODBC application connects to Impala through a load balancer such as `haproxy`, be cautious about reusing the connections. If the load balancer has set up connection timeout values, either check the connection frequently so that it never sits idle longer than the load balancer timeout value, or check the connection validity before using it and create a new one if the connection has been closed.

## Resource Management

Resource management helps ensure predictable behavior by defining the impact of different services on cluster resources. The goals of resource management features are to:

- Guarantee completion in a reasonable time frame for critical workloads
- Support reasonable cluster scheduling between groups of users based on fair allocation of resources per group
- Prevent users from depriving other users access to the cluster

## Schedulers

The scheduler is responsible for deciding which tasks get to run where and when to run them.

The MapReduce and YARN computation frameworks support the following schedulers:

- **FIFO** - Allocates resources based on arrival time.
- **Fair** - Allocates resources to weighted pools, with fair sharing within each pool.
  - [CDH 4 Fair Scheduler](#)
  - [CDH 5 Fair Scheduler](#)
- **Capacity** - Allocates resources to pools, with FIFO scheduling within each pool.
  - [CDH 4 Capacity Scheduler](#)
  - [CDH 5 Capacity Scheduler](#)

The scheduler defaults for MapReduce and YARN are:

- **MapReduce** - Cloudera Manager, CDH 5, and CDH 4 set the default to FIFO. FIFO is set as the default for backward-compatibility purposes, but Cloudera recommends Fair Scheduler because Impala and Llama are optimized for it. Capacity Scheduler is also available.

If you are running CDH 4, you can specify how MapReduce jobs share resources by [configuring the MapReduce scheduler](#).

- **YARN** - Cloudera Manager, CDH 5, and CDH 4 set the default to Fair Scheduler. Cloudera recommends Fair Scheduler because Impala and Llama are optimized for it. FIFO and Capacity Scheduler are also available.

In YARN, the scheduler is responsible for allocating resources to the various running applications subject to familiar constraints of capacities, queues, and so on. The scheduler performs its scheduling function based on the resource requirements of the applications; it does so based on the abstract notion of a resource **container** that incorporates elements such as memory, CPU, disk, network, and so on.

The YARN scheduler has a pluggable policy plug-in, which is responsible for partitioning the cluster resources among the various queues, applications, and so on.

If you are running CDH 5, you can specify how YARN applications share resources by manually [configuring the YARN scheduler](#). Alternatively you can use Cloudera Manager [dynamic allocation](#) features to manage the scheduler configuration.

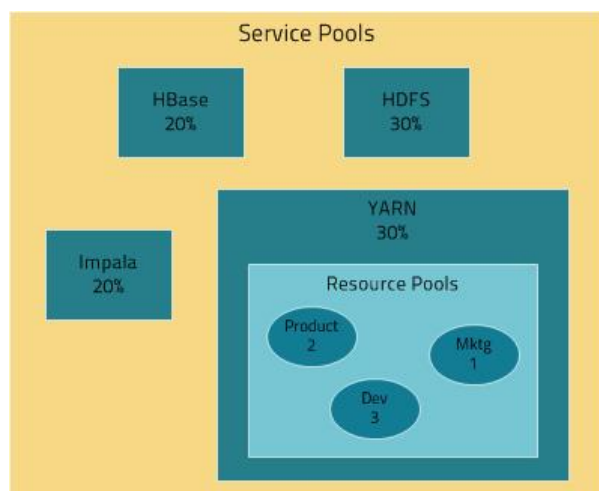
## Cloudera Manager Resource Management Features

Cloudera Manager provides the following features to assist you with allocating cluster resources to services.

### Static Allocation

Cloudera Manager 4 introduced the ability to partition resources across HBase, HDFS, Impala, MapReduce, and YARN services by allowing you to set configuration properties that were enforced by Linux control groups (Linux cgroups). With Cloudera Manager 5, the ability to statically allocate resources using cgroups is configurable through a single *static service pool wizard*. You allocate services a percentage of total resources and the wizard configures the cgroups.

For example, the following figure illustrates static pools for HBase, HDFS, Impala, and YARN services that are respectively assigned 20%, 30%, 20%, and 30% of cluster resources.



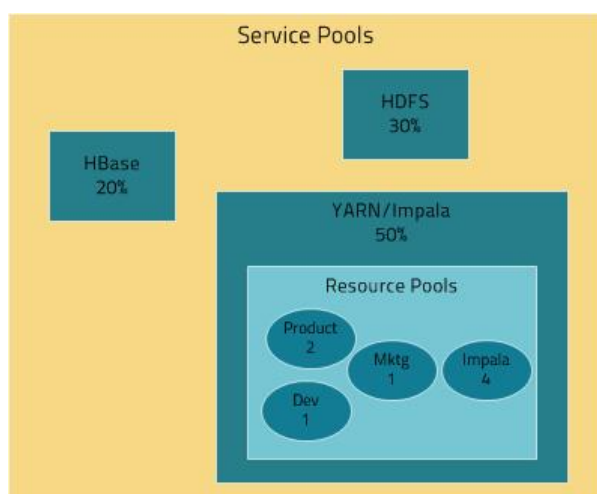
### Dynamic Allocation

Cloudera Manager allows you to manage mechanisms for dynamically apportioning resources statically allocated to YARN and Impala using *dynamic resource pools*.

Depending on the version of CDH you are using, dynamic resource pools in Cloudera Manager support the following resource management (RM) scenarios:

- **(CDH 5) YARN Independent RM** - YARN manages the virtual cores, memory, running applications, and scheduling policy for each pool. In the preceding diagram, three dynamic resource pools - Dev, Product, and Mktg with weights 3, 2, and 1 respectively - are defined for YARN. If an application starts and is assigned to the Product pool, and other applications are using the Dev and Mktg pools, the Product resource pool will receive  $30\% \times \frac{2}{6}$  (or 10%) of the total cluster resources. If there are no applications using the Dev and Mktg pools, the YARN Product pool will be allocated 30% of the cluster resources.
- **(CDH 5) YARN and Impala Integrated RM** - YARN manages memory for pools running Impala queries; Impala limits the number of running and queued queries in each pool. In the YARN and Impala integrated RM scenario, Impala services can reserve resources through YARN, effectively sharing the static YARN service pool and resource pools with YARN applications. The integrated resource management scenario, where both YARN and Impala use the YARN resource management framework, require the [Impala Llama](#) role.

In the following figure, the YARN and Impala services have a 50% static share which is subdivided among the original resource pools with an additional resource pool designated for the Impala service. If YARN applications are using all the original pools, and Impala uses its designated resource pool, Impala queries will have the same resource allocation  $50\% \times \frac{4}{8} = 25\%$  as in the first scenario. However, when YARN applications are not using the original pools, Impala queries will have access to 50% of the cluster resources.



- **(CDH 5) YARN and Impala Independent RM** - YARN manages the virtual cores, memory, running applications, and scheduling policy for each pool; Impala manages memory for pools running queries and limits the number of running and queued queries in each pool.
- **(CDH 5 and CDH 4) Impala Independent RM** - Impala manages memory for pools running queries and limits the number of running and queued queries in each pool.

The scenarios where YARN manages resources, whether for independent RM or integrated RM, map to the YARN scheduler configuration. The scenarios where Impala independently manages resources employ the Impala [admission control](#) feature.

To submit a YARN application to a specific resource pool, specify the `mapreduce.job.queueName` property. The YARN application's queue property is mapped to a resource pool. To submit an Impala query to a specific resource pool, specify the `REQUEST_POOL` option.

## Managing Resources with Cloudera Manager

### Linux Control Groups

[Required Role:](#) **Full Administrator**

Cloudera Manager supports the Linux control groups (cgroups) kernel feature. With cgroups, administrators can impose per-resource restrictions and limits on services and roles. This provides the ability to allocate resources using cgroups to enable isolation of compute frameworks from one another. Resource allocation is implemented by setting properties for the services and roles.

### Linux Distribution Support

Cgroups are a feature of the Linux kernel, and as such, support depends on the host's Linux distribution and version. As shown in the following table, Linux cgroups don't exist on RHEL 5. However, all of the other OS platforms supported by Cloudera Manager support cgroups.

Distribution	CPU Shares	I/O Weight	Memory Soft Limit	Memory Hard Limit
Red Hat Enterprise Linux (or CentOS) 5				
Red Hat Enterprise Linux (or CentOS) 6	■	■	■	■

Distribution	CPU Shares	I/O Weight	Memory Soft Limit	Memory Hard Limit
SUSE Linux Enterprise Server 11	■	■	■	■
Ubuntu 10.04 LTS	■		■	■
Ubuntu 12.04 LTS	■	■	■	■
Ubuntu 12.04 LTS	■	■	■	■
Debian 6.0	■			
Debian 7.0	■			
Debian 7.1	■			

If a distribution lacks support for a given parameter, changes to the parameter have no effect.

The exact level of support can be found in the Cloudera Manager Agent log file, shortly after the Agent has started. See [Viewing Cloudera Manager Server and Agent Logs](#) to find the Agent log. In the log file, look for an entry like this:

```
Found cgroups capabilities: {'has_memory': True, 'default_memory_limit_in_bytes': 9223372036854775807, 'writable_cgroup_dot_procs': True, 'has_cpu': True, 'default_blkio_weight': 1000, 'default_cpu_shares': 1024, 'has_blkio': True}
```

The `has_memory` and similar entries correspond directly to support for the CPU, I/O, and memory parameters.

### Further Reading

- <http://www.kernel.org/doc/Documentation/cgroups/cgroups.txt>
- <http://www.kernel.org/doc/Documentation/cgroups/blkio-controller.txt>
- <http://www.kernel.org/doc/Documentation/cgroups/memory.txt>
- [https://access.redhat.com/site/documentation/en-US/Red\\_Hat\\_Enterprise\\_Linux/6/html/Resource\\_Management\\_Guide/index.html](https://access.redhat.com/site/documentation/en-US/Red_Hat_Enterprise_Linux/6/html/Resource_Management_Guide/index.html)

## Resource Management with Control Groups

In order to use cgroups, you must also enable cgroup-based resource management under the **Host > Resource Management** configuration properties. However, if you configure [static service pools](#), this property will be set as part of that process.

### Enabling Resource Management

Cgroups-based resource management can be enabled for all hosts, or on a per-host basis.

1. If you've upgraded from a version of Cloudera Manager older than Cloudera Manager 4.5, restart every Cloudera Manager Agent before using cgroups-based resource management:

- a. Stop all services, including the Cloudera Management Service.
- b. On each cluster host, run as root:

```
$ sudo service cloudera-scm-agent hard_restart
```

- c. Start all services.

2. Click the **Hosts** tab.
3. Optionally click the link for the host where you want to enable cgroups.
4. Click the **Configuration** tab.
5. Select the **Resource Management** category.
6. Check **Enable Cgroup-based Resource Management** checkbox.

- Restart all roles on the host or hosts.

### Limitations

- Role group and role instance override cgroup-based resource management parameters must be saved one at a time. Otherwise some of the changes that should be reflected dynamically will be ignored.
- The role group abstraction is an imperfect fit for resource management parameters, where the goal is often to take a numeric value for a host resource and distribute it amongst running roles. The role group represents a "horizontal" slice: the same role across a set of hosts. However, the cluster is often viewed in terms of "vertical" slices, each being a combination of slave roles (such as TaskTracker, DataNode, Region Server, Impala Daemon, and so on). Nothing in Cloudera Manager guarantees that these disparate horizontal slices are "aligned" (meaning, that the role assignment is identical across hosts). If they are unaligned, some of the role group values will be incorrect on unaligned hosts. For example a host whose role groups have been configured with memory limits but that's missing a role will probably have unassigned memory.

### Configuring Resource Parameters

After enabling cgroups, you can restrict and limit the resource consumption of roles (or role groups) on a per-resource basis. All of these parameters can be found in the Cloudera Manager Admin Console, under the Resource Management category:

- CPU Shares** - The more CPU shares given to a role, the larger its share of the CPU when under contention. Until processes on the host (including both roles managed by Cloudera Manager and other system processes) are contending for all of the CPUs, this will have no effect. When there is contention, those processes with higher CPU shares will be given more CPU time. The effect is linear: a process with 4 CPU shares will be given roughly twice as much CPU time as a process with 2 CPU shares.

Updates to this parameter will be dynamically reflected in the running role.

- I/O Weight** - The greater the I/O weight, the higher priority will be given to I/O requests made by the role when I/O is under contention (either by roles managed by Cloudera Manager or by other system processes). This only affects read requests; write requests remain unprioritized.

Updates to this parameter will be dynamically reflected in the running role.

- Memory Soft Limit** - When the limit is reached, the kernel will reclaim pages charged to the process if and only if the host is facing memory pressure. If reclaiming fails, the kernel may kill the process. Both anonymous as well as page cache pages contribute to the limit.

After updating this parameter, the role must be restarted before changes take effect.

- Memory Hard Limit** - When a role's resident set size (RSS) exceeds the value of this parameter, the kernel will swap out some of the role's memory. If it's unable to do so, it will kill the process. Note that the kernel measures memory consumption in a manner that doesn't necessarily match what the `top` or `ps` report for RSS, so expect that this limit is a rough approximation.

After updating this parameter, the role must be restarted before changes take effect.

### Example: Protecting Production MapReduce Jobs from Impala Queries

Suppose you have MapReduce deployed in production and want to roll out Impala without affecting production MapReduce jobs. For simplicity, we will make the following assumptions:

- The cluster is using homogenous hardware
- Each worker host has two cores
- Each worker host has 8 GB of RAM
- Each worker host is running a DataNode, TaskTracker, and an Impala Daemon
- Each role type is in a single role group
- Cgroups-based resource management has been enabled on all hosts

Action	Procedure
CPU	<ol style="list-style-type: none"> <li>1. Leave DataNode and TaskTracker role group CPU shares at 1024.</li> <li>2. Set Impala Daemon role group's CPU shares to 256.</li> <li>3. The TaskTracker role group should be configured with a Maximum Number of Simultaneous Map Tasks of 2 and a Maximum Number of Simultaneous Reduce Tasks of 1. This yields an upper bound of three MapReduce tasks at any given time; this is an important detail for memory sizing.</li> </ol>
Memory	<ol style="list-style-type: none"> <li>1. Set Impala Daemon role group memory limit to 1024 MB.</li> <li>2. Leave DataNode maximum Java heap size at 1 GB.</li> <li>3. Leave TaskTracker maximum Java heap size at 1 GB.</li> <li>4. Leave MapReduce Child Java Maximum Heap Size for Gateway at 1 GB.</li> <li>5. Leave cgroups hard memory limits alone. We'll rely on "cooperative" memory limits exclusively, as they yield a nicer user experience than the cgroups-based hard memory limits.</li> </ol>
I/O	<ol style="list-style-type: none"> <li>1. Leave DataNode and TaskTracker role group I/O weight at 500.</li> <li>2. Impala Daemon role group I/O weight is set to 125.</li> </ol>

When you're done with configuration, restart all services for these changes to take effect.

The results are:

1. When MapReduce jobs are running, all Impala queries together will consume up to a fifth of the cluster's CPU resources.
2. Individual Impala Daemons won't consume more than 1 GB of RAM. If this figure is exceeded, new queries will be cancelled.
3. DataNodes and TaskTrackers can consume up to 1 GB of RAM each.
4. We expect up to 3 MapReduce tasks at a given time, each with a maximum heap size of 1 GB of RAM. That's up to 3 GB for MapReduce tasks.
5. The remainder of each host's available RAM (6 GB) is reserved for other host processes.
6. When MapReduce jobs are running, read requests issued by Impala queries will receive a fifth of the priority of either HDFS read requests or MapReduce read requests.

## Static Service Pools

**Static service pools** isolate the services in your cluster from one another, so that load on one service has a bounded impact on other services. Services are allocated a static percentage of total resources—CPU, memory, and I/O weight—which are not shared with other services. When you configure static service pools, Cloudera Manager computes recommended memory, CPU, and I/O configurations for the worker roles of the services that correspond to the percentage assigned to each service. Static service pools are implemented per role group within a cluster, using [Linux control groups \(cgroups\)](#) and cooperative memory limits (for example, Java maximum heap sizes). Static service pools can be used to control access to resources by HBase, HDFS, Impala, MapReduce, Solr, Spark, YARN, and [add-on](#) services. Static service pools are not enabled by default.

### Note:

- I/O allocation only works when [short-circuit reads](#) are enabled.
- I/O allocation does not handle write side I/O because cgroups in the Linux kernel do not currently support buffered writes.



## Viewing Static Service Pool Status

Select **Clusters** > *Cluster name* > **Resource Management** > **Static Service Pools**. If the cluster has a YARN service, the Static Service Pools Status tab displays and shows whether resource management is enabled for the cluster, and the currently configured service pools.

See [Monitoring Static Service Pools](#) for more information.

## Enabling and Configuring Static Service Pools

**Required Role:** **Cluster Administrator** **Full Administrator**

1. Select **Clusters** > *Cluster name* > **Resource Management** > **Static Service Pools**.
2. Click the **Configuration** tab. The **Step 1 of 4: Basic Allocation Setup** page displays. In each field in the basic allocation table, enter the percentage of resources to give to each service. The total must add up to 100%. In CDH 5 clusters, if you [enable integrated resource management](#), the Impala service shares the YARN service pool, rather than use its own static pool. In this case, you cannot specify a percentage for the Impala service. Click **Continue** to proceed.
3. **Step 2: Review Changes** - The allocation of resources for each resource type and role displays with the new values as well as the values previously in effect. The values for each role are set by role group; if there is more than one role group for a given role type (for example, for RegionServers or DataNodes) then resources will be allocated separately for the hosts in each role group. Take note of changed settings. If you have previously customized these settings, check these over carefully.
  - Click the ➤ to the right of each percentage to display the allocations for a single service. Click ➤ to the right of the Total (100%) to view all the allocations in a single page.
  - Click the **Back** button to go to the previous page and change your allocations.

When you are satisfied with the allocations, click **Continue**.

4. **Step 3 of 4: Restart Services** - To apply the new allocation percentages, click **Restart Now** to restart the cluster. To skip this step, click **Restart Later**. If HDFS High Availability is enabled, you will have the option to choose a [rolling restart](#).
5. **Step 4 of 4: Progress** displays the status of the restart commands. Click **Finished** after the restart commands complete.

## Disabling Static Service Pools

**Required Role:** **Cluster Administrator** **Full Administrator**

To disable static service pools, disable cgroup-based resource management for all hosts in all clusters:

1. In the main navigation bar, click **Hosts**.
2. Click the **Configuration** tab.
3. Click the **Resource Management** category, uncheck the **Enable Cgroup-based Resource Management** property, then click **Save Changes**.
4. Restart all services.

Static resource management is disabled, but the percentages you set when you configured the pools, and all the changed settings (for example, heap sizes), *are retained* by the services. The percentages and settings will also be used when you re-enable static service pools. If you want to revert to the settings you had before static service pools were enabled, follow the procedures in [Viewing and Reverting Configuration Changes](#) on page 23.

## Dynamic Resource Pools

**Required Role:** **Configurator** **Cluster Administrator** **Full Administrator**

A **dynamic resource pool** is a named configuration of resources and a policy for scheduling the resources among YARN applications and Impala queries running in the pool. Dynamic resource pools allow you to schedule and

allocate resources to YARN applications and Impala queries based on a user's access to specific pools and the resources available to those pools. If a pool's allocation is not in use it can be given to other pools. Otherwise, a pool receives a share of resources in accordance with the pool's weight. Dynamic resource pools have ACLs that restrict who can submit work to and administer them.

A **configuration set** defines the allocation of resources across pools that may be active at a given time. For example, you can define "weekday" and "weekend" configuration sets, which define different resource pool configurations for different days of the week.

A **scheduling rule** defines when a configuration set is active. The configuration set is updated in affected services every hour.

Resource pools can be nested, with sub-pools restricted by the settings of their parent pool.

The resources available for sharing are subject to the allocations made for each service if [static service pools](#) (cgroups) are being enforced. For example, if the static pool for YARN is 75% of the total cluster resources, then resource pools will use only that 75% of resources.

### Viewing Dynamic Resource Pool Configuration

Depending on which resource management scenario described in [Cloudera Manager Resource Management Features](#) on page 171 is in effect, the dynamic resource pool configuration overview displays the following information:

- **YARN Independent RM** - Weight, Virtual Cores, Min and Max Memory, Max Running Apps, and Scheduling Policy
- **YARN and Impala Integrated RM**
  - **YARN** - Weight, Virtual Cores, Min and Max Memory, Max Running Apps, and Scheduling Policy
  - **Impala** - Max Running Queries and Max Queued Queries
- **YARN and Impala Independent RM**
  - **YARN** - Weight, Virtual Cores, Min and Max Memory, Max Running Apps, and Scheduling Policy
  - **Impala** - Max Memory, Max Running Queries, and Max Queued Queries
- **Impala Independent RM** - Max Memory, Max Running Queries, and Max Queued Queries

To view dynamic resource pool configuration:

1. Select **Clusters** > **Cluster name** > **Resource Management** > **Dynamic Resource Pools**. If the cluster has a YARN service, the Dynamic Resource Pools Status tab displays. If the cluster has only an Impala service enabled for dynamic resource pools, the Dynamic Resource Pools Configuration tab displays.
2. If the Status page is displayed, click the **Configuration** tab. A list of the currently configured pools with their configured limits displays.

### Enabling and Disabling Dynamic Resource Pools for Impala

By default dynamic resource pools for Impala are enabled. If dynamic resource pools are disabled, the Impala section will not appear in the Dynamic Resource Pools tab or in the resource pool dialogs within that page. To modify the Impala dynamic resource pool setting:

1. Go to the Impala service.
2. Click the **Configuration** tab.
3. Expand the **Service-Wide** category.
4. Expand the **Admission Control** category.
5. Click the **Enable Dynamic Resource Pools** property and check or uncheck the checkbox.
6. Click **Save Changes** to commit the changes.
7. Restart the Impala service.


### Creating a Dynamic Resource Pool

1. Select **Clusters > Cluster name > Resource Management > Dynamic Resource Pools**. If the cluster has a YARN service, the Dynamic Resource Pools Status tab displays. If the cluster has only an Impala service enabled for dynamic resource pools, the Dynamic Resource Pools Configuration tab displays.
2. If the Status page is displayed, click the **Configuration** tab. A list of the currently configured pools with their configured limits displays.
3. Click **Add Resource Pool**. The Add dialog box displays showing the **General** tab.
4. Specify a name and resource limits for the pool:
  - In the **Resource Pool Name** field, type a unique name containing only alphanumeric characters.
  - Specify the policy for scheduling resource among applications running in the pool:
    - **Dominant Resource Fairness (DRF) (default)** - An extension of fair scheduling for more than one resource—it determines resource shares (CPU, memory) for a job separately based on the availability of those resources and the needs of the job.
    - **Fair Scheduler (FAIR)** - Determines resource shares based on memory.
    - **First-In, First-Out (FIFO)** - Determines resource shares based on when the job was added.
  - If you have enabled Fair Scheduler preemption, optionally set a preemption timeout to specify how long a job in this pool must wait before it can preempt resources from jobs in other pools. To enable preemption, click the [Fair Scheduler Preemption](#) link or follow the procedure in [Enabling Preemption](#) on page 181.
5. Do one or more of the following:
  - Click the **YARN** tab.
    1. Click a [configuration set](#).
    2. Specify a weight that indicates that pool's share of resources relative to other pools, minimum and maximums for virtual cores and memory, and a limit on the number of applications that can run simultaneously in this pool.
  - Click the **Impala** tab.
    1. Click a [configuration set](#).
    2. Specify the maximum number of concurrently running and queued queries in this pool.
6. If you have [enabled ACLs and specified users or groups](#), optionally click the **Submission** and **Administration Access Control** tabs to specify which users and groups can have submission authorization and can submit applications and which users have administration authorization and can view all and kill applications. The default is that anyone can both submit, view all, and kill applications. To restrict either of these permissions, select the **Allow these users and groups** radio button and provide a comma-delimited list of users and groups in the Users and Groups fields respectively. Click **OK**.

### Configuring Default YARN Scheduler Properties

1. Select **Clusters > Cluster name > Resource Management > Dynamic Resource Pools**. If the cluster has a YARN service, the Dynamic Resource Pools Status tab displays. If the cluster has only an Impala service enabled for dynamic resource pools, the Dynamic Resource Pools Configuration tab displays.
2. If the Status page is displayed, click the **Configuration** tab. A list of the currently configured pools with their configured limits displays.
3. Click the **Default Settings** button.
4. Specify the default scheduling policy, maximum applications, and preemption timeout properties.
5. Click **OK**.

### Editing Dynamic Resource Pools

After you edit a resource pool configuration, the  **Refreshing** displays while the settings are propagated to the scheduler configuration file.

1. Select **Clusters** > **Cluster name** > **Resource Management** > **Dynamic Resource Pools**. If the cluster has a YARN service, the Dynamic Resource Pools Status tab displays. If the cluster has only an Impala service enabled for dynamic resource pools, the Dynamic Resource Pools Configuration tab displays.
2. If the Status page is displayed, click the **Configuration** tab. A list of the currently configured pools with their configured limits displays.
3. Click **Edit** at the right of a resource pool row. Edit the properties and click **OK**.
4. If you have [enabled ACLs and specified users or groups](#), optionally click the **Submission** and **Administration Access Control** tabs to specify which users and groups can have submission authorization and can submit applications and which users have administration authorization and can view all and kill applications. The default is that anyone can both submit, view all, and kill applications. To restrict either of these permissions, select the **Allow these users and groups** radio button and provide a comma-delimited list of users and groups in the Users and Groups fields respectively. Click **OK**.

### Adding Sub-Pools

Pools can be nested as sub-pools. They share among their siblings the resources of the parent pool. Each sub-pool can have its own resource restrictions; if those restrictions fall within the configuration of the parent pool, then the limits for the sub-pool take effect. If the limits for the sub-pool exceed those of the parent, then the parent limits take effect.

Once you create sub-pools, jobs cannot be submitted to the parent pool; they must be submitted to a sub-pool.

1. Select **Clusters** > **Cluster name** > **Resource Management** > **Dynamic Resource Pools**. If the cluster has a YARN service, the Dynamic Resource Pools Status tab displays. If the cluster has only an Impala service enabled for dynamic resource pools, the Dynamic Resource Pools Configuration tab displays.
2. If the Status page is displayed, click the **Configuration** tab. A list of the currently configured pools with their configured limits displays.
3. Click ▼ at the right of a resource pool row and select **Add Sub Pool**. Configure sub-pool properties.
4. Click **OK**.

## YARN Pool Status and Configuration Options

### Viewing Dynamic Resource Pool Status

Select **Clusters** > **ClusterName** > **Dynamic Resource Pools**. The Status tab displays the YARN resource pools currently in use for the cluster. See [Monitoring Dynamic Resource Pools](#) for more information.



### Setting User Limits

Pool properties determine the maximum number of applications that can run in a pool. To limit the number of applications specific users can run at the same time in a pool:

1. Select **Clusters** > **Cluster name** > **Resource Management** > **Dynamic Resource Pools**.
2. Click the **Configuration** tab.
3. Click the **User Limits** tab. The table displays a list of users and the maximum number of jobs each user can submit.
4. Click **Add User Limit**.
5. Specify a username and maximum number of running applications.
6. Click **OK**.



### Enabling ACLs

To specify whether ACLs are checked:

1. Select **Clusters** > *Cluster name* > **Resource Management** > **Dynamic Resource Pools**.
2. Click the **Configuration** tab.
3. Click **Other Settings**.
4. In the **Enable ResourceManager ACLs** property, click . The YARN service configuration page displays.
5. Select the checkbox.
6. Click **Save Changes** to commit the changes.
7. Click  to invoke the cluster restart wizard.
8. Click **Restart Cluster**.
9. Click **Restart Now**.
10. Click **Finish**.



### Configuring ACLs

To configure which users and groups can submit and kill YARN applications in any resource pool:

1. [Enable ACLs](#).
2. Select **Clusters** > *Cluster name* > **Resource Management** > **Dynamic Resource Pools**.
3. Click the **Configuration** tab.
4. Click **Other Settings**.
5. In the **Admin ACL** property, click . The YARN service configuration page displays.
6. Specify which users and groups can submit and kill applications.
7. Click **Save Changes** to commit the changes.
8. Click  to invoke the cluster restart wizard.
9. Click **Restart Cluster**.
10. Click **Restart Now**.
11. Click **Finish**.

### Enabling Preemption

You can enable the Fair Scheduler to preempt applications in other pools if a pool's minimum share is not met for some period of time. When you [create a pool](#) you can specify how long a pool must wait before other applications are preempted.

1. Select **Clusters** > *Cluster name* > **Resource Management** > **Dynamic Resource Pools**.
2. Click the **Configuration** tab.
3. Click the **User Limits** tab. The table shows you a list of users and the maximum number of jobs each user can submit.
4. Click **Other Settings**.
5. In the **Fair Scheduler Preemption**, click . The YARN service configuration page displays.
6. Select the checkbox.
7. Click **Save Changes** to commit the changes.
8. Click  to invoke the cluster restart wizard.
9. Click **Restart Cluster**.
10. Click **Restart Now**.
11. Click **Finish**.

### Placement Rules


Cloudera Manager provides many options for determining how YARN applications and Impala queries are placed in resource pools. You can specify basic rules that place applications and queries in pools based on runtime

configurations or the name of the user running the application or query or select an advanced option that allows you to specify a set of ordered rules for placing applications and queries in pools.

To submit a YARN application to a specific resource pool, specify the `mapreduce.job.queue.name` property. The YARN application's queue property is mapped to a resource pool. To submit an Impala query to a specific resource pool, specify the `REQUEST_POOL` option.

### Enabling and Disabling Undeclared Pools


If you do not specify a pool with a job or query property, by default YARN and Impala create a pool "on-the-fly" with the name of the user that submitted the request and assigns it to that resource pool. For YARN, you can change this behavior so that the **default** pool is used instead:

1. Select **Clusters > Cluster name > Resource Management > Dynamic Resource Pools**. If the cluster has a YARN service, the Dynamic Resource Pools Status tab displays. If the cluster has only an Impala service enabled for dynamic resource pools, the Dynamic Resource Pools Configuration tab displays.
2. If the Status page is displayed, click the **Configuration** tab. A list of the currently configured pools with their configured limits displays.
3. Click the **Placement Rules** tab.
4. Click **Basic** radio button.
5. Click the **Allow Undeclared Pools** property.
6. Select or deselect the **Allow Undeclared Pools** checkbox.
7. Click **Save Changes** to commit the changes.
8. Click  to invoke the cluster restart wizard.
9. Click **Restart Cluster**.
10. Click **Restart Now**.
11. Click **Finish**.

■ **Note:** YARN and Impala pools created "on-the-fly" are deleted when you restart the YARN and Impala services.

### Enabling and Disabling the Default Pool

If an application specifies a pool that has not been explicitly configured or is assigned to a pool with the name of user according to the **Fair Scheduler User As Default Queue** property, by default YARN creates the pool at runtime with default settings. To change the behavior so that under these circumstances the **default** pool is used instead:

1. Select **Clusters > Cluster name > Resource Management > Dynamic Resource Pools**. If the cluster has a YARN service, the Dynamic Resource Pools Status tab displays. If the cluster has only an Impala service enabled for dynamic resource pools, the Dynamic Resource Pools Configuration tab displays.
2. If the Status page is displayed, click the **Configuration** tab. A list of the currently configured pools with their configured limits displays.
3. Click the **Placement Rules** tab.
4. Click **Basic** radio button.
5. Click the **Fair Scheduler User As Default Queue** property.
6. Select or deselect the checkbox.
7. Click **Save Changes** to commit the changes.
8. Click  to invoke the cluster restart wizard.
9. Click **Restart Cluster**.
10. Click **Restart Now**.
11. Click **Finish**.

## Specifying Advanced Placement Rules and Rule Ordering

You use **placement rules** to indicate whether applications are placed in specified pools, pools named by a user or group, or the default pool. To configure and order a set of rules:

1. Select the **Advanced** radio button on the Placement Rules tab.
2. Click **+** to add a new rule row and **-** to remove a rule row.
3. In each row, click **▼** and select a rule. The available rules are:
  - specified pool; create the pool if it doesn't exist (default 1st)
  - root.<username> pool; create the pool if it doesn't exist (default 2nd) - the application or query is placed into a pool with the name of the user who submitted it.
  - specified pool only if the pool exists
  - root.<username> pool only if the pool exists
  - root.<primaryGroupName> pool; create the pool if it doesn't exist - the application or query is placed into a pool with the name of the primary group of the user who submitted it.
  - root.<primaryGroupName> pool only if the pool exists
  - root.<secondaryGroupName> pool only if one of these pools exists - the application or query is placed into a pool with a name that matches a secondary group of the user who submitted it.
  - default pool; create the pool if it doesn't exist

For more information about these rules, see the description of the `queuePlacementPolicy` element in [Allocation File Format](#). Reorder rules by selecting different rules for existing rule rows. If a rule is always satisfied, subsequent rules are not evaluated and appear disabled.

4. Click **Save**. The [Fair Scheduler](#) allocation file (by default, `fair-scheduler.xml`) is updated.

## Configuration Sets

A **configuration set** defines the allocation of resources across pools that may be active at a given time. For example, you can define "weekday" and "weekend" configuration sets, which define different resource pool configurations for different days of the week.

### Creating a Configuration Set

1. Select **Clusters > Cluster name > Resource Management > Dynamic Resource Pools**. If the cluster has a YARN service, the Dynamic Resource Pools Status tab displays. If the cluster has only an Impala service enabled for dynamic resource pools, the Dynamic Resource Pools Configuration tab displays.
2. If the Status page is displayed, click the **Configuration** tab. A list of the currently configured pools with their configured limits displays.
3. Click the **Scheduling Rules** tab.
4. Click **Add Scheduling Rule**.
5. In the Configuration Set field, select the **Create New** radio button.
- 6.
7. Click **Add Configuration Set**. The Add Configuration Set dialog displays.
  - a. Type a name in the Name field and select the configuration set to clone from the **Clone from Configuration Set** drop-down.
  - b. Click **OK**. The configuration set is added to and selected in the **Configuration Sets** drop-down.
8. For each resource pool, click **Edit**.
  - a. Select a resource pool configuration set name.
  - b. Edit the pool properties and click **OK**.
9. Define one or more [scheduling rules](#) to specify when the configuration set is active.

### Example Configuration Sets

The **weekday** configuration set assigns the **production** pool four times the resources of the **development** pool:

+ Add Resource Pool		🔧 Default Settings		Configuration Sets				weekday ▾		
Name	YARN						Impala			
	Weight	%	Virtual Cores Min / Max	Memory Min / Max	Max Running Apps	Scheduling Policy	Max Running Queries	Max Queued Queries		
root	1	100.0%	- / -	- / -	-	DRF	-	-	<a href="#">🔗 Edit</a>	▾
production	4	80.0%	- / -	- / -	-	DRF	-	-	<a href="#">🔗 Edit</a>	▾
development	1	20.0%	- / -	- / -	-	DRF	-	-	<a href="#">🔗 Edit</a>	▾

The **weekend** configuration set assigns the **production** and **development** pools an equal share of the resources:

+ Add Resource Pool		🔧 Default Settings		Configuration Sets				weekend ▾		
Name	YARN						Impala			
	Weight	%	Virtual Cores Min / Max	Memory Min / Max	Max Running Apps	Scheduling Policy	Max Running Queries	Max Queued Queries		
root	1	100.0%	- / -	- / -	-	DRF	-	-	<a href="#">🔗 Edit</a>	▾
production	1	50.0%	- / -	- / -	-	DRF	-	-	<a href="#">🔗 Edit</a>	▾
development	1	50.0%	- / -	- / -	-	DRF	-	-	<a href="#">🔗 Edit</a>	▾

The **default** configuration set assigns the **production** pool twice the resources of the **development** pool:

+ Add Resource Pool

🔧 Default Settings

Configuration Sets

default ▾

Name	YARN						Impala		
	Weight	%	Virtual Cores Min / Max	Memory Min / Max	Max Running Apps	Scheduling Policy	Max Running Queries	Max Queued Queries	
root	1	100.0%	- / -	- / -	-	DRF	-	-	🔗 Edit ▾
production	2	66.7%	- / -	- / -	-	DRF	-	-	🔗 Edit ▾
development	1	33.3%	- / -	- / -	-	DRF	-	-	🔗 Edit ▾

See [example scheduling rules](#) for these configuration sets.

## Viewing the Properties of a Configuration Set

1. Select **Clusters** > **Cluster name** > **Resource Management** > **Dynamic Resource Pools**. If the cluster has a YARN service, the Dynamic Resource Pools Status tab displays. If the cluster has only an Impala service enabled for dynamic resource pools, the Dynamic Resource Pools Configuration tab displays.
2. If the Status page is displayed, click the **Configuration** tab. A list of the currently configured pools with their configured limits displays.
3. In the **Configuration Sets** drop-down, select a configuration set. The properties of each pool for that configuration set display.

## Deleting a Configuration Set

1. Select **Clusters** > **Cluster name** > **Resource Management** > **Dynamic Resource Pools**. If the cluster has a YARN service, the Dynamic Resource Pools Status tab displays. If the cluster has only an Impala service enabled for dynamic resource pools, the Dynamic Resource Pools Configuration tab displays.



2. If the Status page is displayed, click the **Configuration** tab. A list of the currently configured pools with their configured limits displays.
3. In the **Configuration Sets** drop-down, select a configuration set. The properties of each pool for that configuration set display.
4. Click **Delete**.

## Scheduling Rules

A **scheduling rule** defines when a configuration set is active. The configuration set is updated in affected services every hour.

### Example Scheduling Rules

Consider the [example weekday and weekend](#) configuration sets. To specify that the **weekday** configuration set is active every weekday, **weekend** configuration set is active on the weekend (weekly on Saturday and Sunday), and the **default** configuration set is active all other times, define the following rules:

+ Add Scheduling Rule   ✕ Reorder Scheduling Rules	
Scheduling Rule	Configuration Set
Repeats weekly on Monday, Tuesday, Wednesday, Thursday, Friday from 12:00 AM to 12:00 AM (PDT), starting 03/24/2014.	<b>weekday</b> <a href="#">Edit</a>
Repeats weekly on Sunday, Saturday from 12:00 AM to 12:00 AM (PDT), starting 03/24/2014.	<b>weekend</b> <a href="#">Edit</a>
Runs when all other rules don't apply.	<b>default</b>

### Adding a Scheduling Rule

1. Select **Clusters > Cluster name > Resource Management > Dynamic Resource Pools**. If the cluster has a YARN service, the Dynamic Resource Pools Status tab displays. If the cluster has only an Impala service enabled for dynamic resource pools, the Dynamic Resource Pools Configuration tab displays.
2. If the Status page is displayed, click the **Configuration** tab. A list of the currently configured pools with their configured limits displays.
3. Click the **Scheduling Rules** tab.
4. Click **Add Scheduling Rule**.
5. In the **Configuration Set** drop-down, select a configuration set.
6. Choose whether the rule should repeat, the repeat frequency, and if the frequency is weekly, the repeat day or days.
7. If the schedule is not repeating, click the left side of the **on** field to display a drop-down calendar where you set the starting date and time. When you specify the date and time, a default time window of two hours is set in the right side of the **on** field. Click the right side to adjust the date and time.
8. Click **OK**.

### Editing a Scheduling Rule

1. Select **Clusters > Cluster name > Resource Management > Dynamic Resource Pools**. If the cluster has a YARN service, the Dynamic Resource Pools Status tab displays. If the cluster has only an Impala service enabled for dynamic resource pools, the Dynamic Resource Pools Configuration tab displays.
2. If the Status page is displayed, click the **Configuration** tab. A list of the currently configured pools with their configured limits displays.
3. Click **Scheduling Rules**.
4. Click **Edit** at the right of a rule.
5. Edit the rule as desired.
6. Click **OK**.

### Deleting a Scheduling Rule

1. Select **Clusters** > **Cluster name** > **Resource Management** > **Dynamic Resource Pools**. If the cluster has a YARN service, the Dynamic Resource Pools Status tab displays. If the cluster has only an Impala service enabled for dynamic resource pools, the Dynamic Resource Pools Configuration tab displays.
2. If the Status page is displayed, click the **Configuration** tab. A list of the currently configured pools with their configured limits displays.
3. Click **Scheduling Rules**.
4. Click ▼ at the right of a rule and select **Delete**.
5. Click **OK**.

## Managing Impala Admission Control

Required Role: **Configurator** **Cluster Administrator** **Full Administrator**

Admission control is an Impala feature that imposes limits on concurrent SQL queries, to avoid resource usage spikes and out-of-memory conditions on busy CDH clusters. It is a form of “throttling”. New queries are accepted and executed until certain conditions are met, such as too many queries or too much total memory used across the cluster. When one of these thresholds is reached, incoming queries wait to begin execution. These queries are queued and are admitted (that is, begin executing) when the resources become available.

For further information on Impala admission control, see [Admission Control and Query Queuing](#) on page 188.

### Enabling and Disabling Impala Admission Control Using Cloudera Manager

1. Go to the Impala service.
2. Click the **Configuration** tab.
3. Expand the **Service-Wide** category.
4. Click the **Admission Control** subcategory.
5. Click the **Enable Impala Admission Control** property and check or uncheck the checkbox.
6. Click **Save Changes** to commit the changes.
7. Restart the Impala service.

### Configuring Impala Admission Control Using Cloudera Manager

1. Go to the Impala service.
2. Click the **Configuration** tab.
3. Expand the **Service-Wide** category.
4. Click the **Admission Control** subcategory. Configure the properties described in [Admission Control Options](#) on page 191.
5. Click **Save Changes** to commit the changes.
6. Restart the Impala service.

## Managing the Impala Llama ApplicationMaster

The Impala Llama ApplicationMaster (Llama) reserves and releases YARN-managed resources for Impala, thus reducing resource management overhead when performing Impala queries. Llama is used when you want to enable integrated resource management.

By default, YARN allocates resources bit-by-bit as needed by MapReduce jobs. Impala needs all resources available at the same time, so that intermediate results can be exchanged between cluster nodes, and queries do not stall partway through waiting for new resources to be allocated. Llama is the intermediary process that ensures all requested resources are available before each Impala query actually begins.

For more information about Llama, see [Llama - Low Latency Application MAster](#).

For information on enabling Llama high availability, see [Llama High Availability](#) on page 260.

## Enabling Integrated Resource Management Using Cloudera Manager

**Required Role:** **Cluster Administrator** **Full Administrator**

The Enable Integrated Resource Management wizard enables cgroups for the *all the hosts* in the cluster running Impala and YARN, adds one or more Llama roles to the Impala service, and configures the Impala and YARN services.

1. Start the wizard using one of the following paths:
  - Cluster-level
    1. Select **Clusters** > **ClusterName** > **Dynamic Resource Pools**.
    2. In the Status section, click **Enable**.
  - Service-level
    1. Go to the Impala service.
    2. Select **Actions** > **Enable Integrated Resource Management**.

The Enable Integrated Resource Management wizard starts and displays information about resource management options and the actions performed by the wizard.

2. Click **Continue**.
3. Leave the **Enable Cgroup-based Resource Management** checkbox checked and click **Continue**.
4. Click the **Impala Llama ApplicationMaster Hosts** field to display a dialog for choosing Llama hosts.

The following shortcuts for specifying hostname patterns are supported:

- Range of hostnames (without the domain portion)

Range Definition	Matching Hosts
10.1.1.[1-4]	10.1.1.1, 10.1.1.2, 10.1.1.3, 10.1.1.4
host[1-3].company.com	host1.company.com, host2.company.com, host3.company.com
host[07-10].company.com	host07.company.com, host08.company.com, host09.company.com, host10.company.com

- IP addresses
  - Rack name
5. Specify or select one or more hosts and click **OK**.
  6. Click **Continue**. A progress screen displays with a summary of the wizard actions.
  7. Click **Continue**.
  8. Click **Restart Now** to restart the cluster and apply the configuration changes or click **leave this wizard** to restart at a later time.
  9. Click **Finish**.

## Disabling Integrated Resource Management Using Cloudera Manager

**Required Role:** **Cluster Administrator** **Full Administrator**

The Enable Integrated Resource Management wizard enables cgroups for the *all the hosts* in the cluster running Impala and YARN, adds one or more Llama roles to the Impala service, and configures the Impala and YARN services.

1. Start the wizard using one of the following paths:
  - 1. Select **Clusters** > **ClusterName** > **Dynamic Resource Pools**.
  - 2. In the Status section, click **Disable**.

## Resource Management

- 1. Go to the Impala service.
- 2. Select **Actions** > **Disable Integrated Resource Management**.

The Disable Integrated Resource Management wizard starts and displays information about resource management options and the actions performed by the wizard.

- 2. Click **Finish**. Integrated resource management is disabled, but resource management using cgroups is left enabled.

## Configuring Llama Using Cloudera Manager

Required Role: **Configurator** **Cluster Administrator** **Full Administrator**

1. Go to the Impala service.
2. Click the **Configuration** tab.
3. Expand the **Impala Llama ApplicationMaster Default Group** subcategory.
4. Click the **Advanced** subcategory.
5. Edit [configuration properties](#).
6. Click **Save Changes** to commit the changes.
7. Restart the Llama role.

## Impala Resource Management

Impala supports two types of resource management: independent and integrated. Independent resource management is supported for CDH 4 and CDH 5 and is implemented by admission control. Integrated resource management is supported for CDH 5 and is implemented by YARN and Llama.

### Admission Control and Query Queuing

Admission control is an Impala feature that imposes limits on concurrent SQL queries, to avoid resource usage spikes and out-of-memory conditions on busy CDH clusters. It is a form of “throttling”. New queries are accepted and executed until certain conditions are met, such as too many queries or too much total memory used across the cluster. When one of these thresholds is reached, incoming queries wait to begin execution. These queries are queued and are admitted (that is, begin executing) when the resources become available.

In addition to the threshold values for currently executing queries, you can place limits on the maximum number of queries that are queued (waiting) and a limit on the amount of time they might wait before returning with an error. These queue settings let you ensure that queries do not wait indefinitely, so that you can detect and correct “starvation” scenarios.

Enable this feature if your cluster is underutilized at some times and overutilized at others. Overutilization is indicated by performance bottlenecks and queries being cancelled due to out-of-memory conditions, when those same queries are successful and perform well during times with less concurrent load. Admission control works as a safeguard to avoid out-of-memory conditions during heavy concurrent usage.

■ **Important:**

- Cloudera strongly recommends you upgrade to CDH 5 or higher to use admission control. In CDH 4, admission control will only work if you *don't* have Hue deployed; unclosed Hue queries will accumulate and exceed the queue size limit. On CDH 4, to use admission control, you must explicitly enable it by specifying `--disable_admission_control=false` in the `impalad` command-line options.
- Use the `COMPUTE STATS` statement for large tables involved in join queries, and follow other steps from [Tuning Impala for Performance](#) to tune your queries. Although `COMPUTE STATS` is an important statement to help optimize query performance, it is especially important when admission control is enabled:
  - When queries complete quickly and are tuned for optimal memory usage, there is less chance of performance or capacity problems during times of heavy load.
  - The admission control feature also relies on the statistics produced by the `COMPUTE STATS` statement to generate accurate estimates of memory usage for complex queries. If the estimates are inaccurate due to missing statistics, Impala might hold back queries unnecessarily even though there is sufficient memory to run them, or might allow queries to run that end up exceeding the memory limit and being cancelled.

## Overview of Impala Admission Control

On a busy CDH cluster, you might find there is an optimal number of Impala queries that run concurrently. Because Impala queries are typically I/O-intensive, you might not find any throughput benefit in running more concurrent queries when the I/O capacity is fully utilized. Because Impala by default cancels queries that exceed the specified memory limit, running multiple large-scale queries at once can result in having to re-run some queries that are cancelled.

The admission control feature lets you set a cluster-wide upper limit on the number of concurrent Impala queries and on the memory used by those queries. Any additional queries are queued until the earlier ones finish, rather than being cancelled or running slowly and causing contention. As other queries finish, the queued queries are allowed to proceed.

For details on the internal workings of admission control, see [How Impala Schedules and Enforces Limits on Concurrent Queries](#) on page 190.

## How Impala Admission Control Relates to YARN

The admission control feature is similar in some ways to the YARN resource management framework, and they can be used separately or together. This section describes some similarities and differences, to help you decide when to use one, the other, or both together.

Admission control is a lightweight, decentralized system that is suitable for workloads consisting primarily of Impala queries and other SQL statements. It sets “soft” limits that smooth out Impala memory usage during times of heavy load, rather than taking an all-or-nothing approach that cancels jobs that are too resource-intensive.

Because the admission control system is not aware of other Hadoop workloads such as MapReduce jobs, you might use YARN with static service pools on heterogeneous CDH 5 clusters where resources are shared between Impala and other Hadoop components. Devote a percentage of cluster resources to Impala, allocate another percentage for MapReduce and other batch-style workloads; let admission control handle the concurrency and memory usage for the Impala work within the cluster, and let YARN manage the remainder of work within the cluster.

You could also try out the combination of YARN, Impala, and Llama, where YARN manages all cluster resources and Impala queries request resources from YARN by using the Llama component as an intermediary. YARN is a more centralized, general-purpose service, with somewhat higher latency than admission control due to the requirement to pass requests back and forth through the YARN and Llama components.

The Impala admission control feature uses the same mechanism as the YARN resource manager to map users to pools and authenticate them. Although the YARN resource manager is only available with CDH 5 and higher, internally Impala includes the necessary infrastructure to work consistently on both CDH 4 and CDH 5. You do not need to run the YARN and Llama components for admission control to operate.

In Cloudera Manager, the controls for Impala resource management change slightly depending on whether the Llama role is enabled, which brings Impala under the control of YARN. When you use Impala without the Llama role, you can specify three properties (memory limit, query queue size, and queue timeout) for the admission control feature. When the Llama role is enabled, you can specify query queue size and queue timeout, but the memory limit is enforced by YARN and not settable through resource pools.

### How Impala Schedules and Enforces Limits on Concurrent Queries

The admission control system is decentralized, embedded in each Impala daemon and communicating through the statestore mechanism. Although the limits you set for memory usage and number of concurrent queries apply cluster-wide, each Impala daemon makes its own decisions about whether to allow each query to run immediately or to queue it for a less-busy time. These decisions are fast, meaning the admission control mechanism is low-overhead, but might be imprecise during times of heavy load. There could be times when the query queue contained more queries than the specified limit, or when the estimated of memory usage for a query is not exact and the overall memory usage exceeds the specified limit. Thus, you typically err on the high side for the size of the queue, because there is not a big penalty for having a large number of queued queries; and you typically err on the low side for the memory limit, to leave some headroom for queries to use more memory than expected, without being cancelled as a result.

At any time, the set of queued queries could include queries submitted through multiple different Impala daemon hosts. All the queries submitted through a particular host will be executed in order, so a `CREATE TABLE` followed by an `INSERT` on the same table would succeed. Queries submitted through different hosts are not guaranteed to be executed in the order they were received. Therefore, if you are using load-balancing or other round-robin scheduling where different statements are submitted through different hosts, set up all table structures ahead of time so that the statements controlled by the queuing system are primarily queries, where order is not significant. Or, if a sequence of statements needs to happen in strict order (such as an `INSERT` followed by a `SELECT`), submit all those statements through a single session, while connected to the same Impala daemon host.

The limit on the number of concurrent queries is a “soft” one. To achieve high throughput, Impala makes quick decisions at the host level about which queued queries to dispatch. Therefore, Impala might slightly exceed the limit from time to time.

To avoid a large backlog of queued requests, you can also set an upper limit on the size of the queue for queries that are delayed. When the number of queued queries exceeds this limit, further queries are cancelled rather than being queued. You can also configure a timeout period, after which queued queries are cancelled, to avoid indefinite waits. If a cluster reaches this state where queries are cancelled due to too many concurrent requests or long waits for query execution to begin, that is a signal for an administrator to take action, either by provisioning more resources, scheduling work on the cluster to smooth out the load, or by doing [Impala performance tuning](#) to enable higher throughput.

### How Admission Control works with Impala Clients (JDBC, ODBC, HiveServer2)

Most aspects of admission control work transparently with client interfaces such as JDBC and ODBC:

- If a SQL statement is put into a queue rather than running immediately, the API call blocks until the statement is dequeued and begins execution. At that point, the client program can request to fetch results, which might also block until results become available.
- If a SQL statement is cancelled because it has been queued for too long or because it exceeded the memory limit during execution, the error is returned to the client program with a descriptive error message.

If you want to submit queries to different resource pools through the `REQUEST_POOL` query option, as described in [REQUEST\\_POOL Query Option](#), In Impala 2.0 and higher you can change that query option through a SQL `SET` statement that you submit from the client application, in the same session. Prior to Impala 2.0, that option was

only settable for a session through the `impala-shell SET` command, or cluster-wide through an `impalad` startup option.

Admission control has the following limitations or special behavior when used with JDBC or ODBC applications:

- The `MEM_LIMIT` query option, sometimes useful to work around problems caused by inaccurate memory estimates for complicated queries, is only settable through the `impala-shell` interpreter and cannot be used directly through JDBC or ODBC applications.
- Admission control does not use the other resource-related query options, `RESERVATION_REQUEST_TIMEOUT` or `V_CPU_CORES`. Those query options only apply to the YARN resource management framework.

## Configuring Admission Control

The configuration options for admission control range from the simple (a single resource pool with a single set of options) to the complex (multiple resource pools with different options, each pool handling queries for a different set of users and groups). You can configure the settings through the Cloudera Manager user interface, or on a system without Cloudera Manager by editing configuration files or through startup options to the `impalad` daemon.

### Admission Control Options

The following Impala configuration options let you adjust the settings of the admission control feature. When supplying the options on the command line, prepend the option name with `--`.

#### `default_pool_max_queued`

**Purpose:** Maximum number of requests allowed to be queued before rejecting requests. Because this limit applies cluster-wide, but each Impala node makes independent decisions to run queries immediately or queue them, it is a soft limit; the overall number of queued queries might be slightly higher during times of heavy load. A negative value or 0 indicates requests are always rejected once the maximum concurrent requests are executing. Ignored if `fair_scheduler_config_path` and `llama_site_path` are set.

**Type:** `int64`

**Default:** 200

#### `default_pool_max_requests`

**Purpose:** Maximum number of concurrent outstanding requests allowed to run before incoming requests are queued. Because this limit applies cluster-wide, but each Impala node makes independent decisions to run queries immediately or queue them, it is a soft limit; the overall number of concurrent queries might be slightly higher during times of heavy load. A negative value indicates no limit. Ignored if `fair_scheduler_config_path` and `llama_site_path` are set.

**Type:** `int64`

**Default:** 200

#### `default_pool_mem_limit`

**Purpose:** Maximum amount of memory (across the entire cluster) that all outstanding requests in this pool can use before new requests to this pool are queued. Specified in bytes, megabytes, or gigabytes by a number followed by the suffix `B` (optional), `m`, or `g`, either uppercase or lowercase. You can specify floating-point values for megabytes and gigabytes, to represent fractional numbers such as `1.5`. You can also specify it as a percentage of the physical memory by specifying the suffix `%`. 0 or no setting indicates no limit. Defaults to bytes if no unit is given. Because this limit applies cluster-wide, but each Impala node makes independent decisions to run queries immediately or queue them, it is a soft limit; the overall memory used by concurrent queries might be slightly higher during times of heavy load. Ignored if `fair_scheduler_config_path` and `llama_site_path` are set.

- **Note:** Impala relies on the statistics produced by the `COMPUTE STATS` statement to estimate memory usage for each query. See [COMPUTE STATS Statement](#) for guidelines about how and when to use this statement.

**Type:** string

**Default:** "" (empty string, meaning unlimited)

`disable_admission_control`

**Purpose:** Turns off the admission control feature entirely, regardless of other configuration option settings.

**Type:** Boolean

**Default:** true

`disable_pool_max_requests`

**Purpose:** Disables all per-pool limits on the maximum number of running requests.

**Type:** Boolean

**Default:** false

`disable_pool_mem_limits`

**Purpose:** Disables all per-pool mem limits.

**Type:** Boolean

**Default:** false

`fair_scheduler_allocation_path`

**Purpose:** Path to the fair scheduler allocation file (`fair-scheduler.xml`).

**Type:** string

**Default:** "" (empty string)

**Usage notes:** Admission control only uses a small subset of the settings that can go in this file, as described below. For details about all the Fair Scheduler configuration settings, see the [Apache wiki](#).

`llama_site_path`

**Purpose:** Path to the Llama configuration file (`llama-site.xml`). If set, `fair_scheduler_allocation_path` must also be set.

**Type:** string

**Default:** "" (empty string)

**Usage notes:** Admission control only uses a small subset of the settings that can go in this file, as described below. For details about all the Llama configuration settings, see the [documentation on Github](#).

`queue_wait_timeout_ms`

**Purpose:** Maximum amount of time (in milliseconds) that a request waits to be admitted before timing out.

**Type:** int64

**Default:** 60000

### Configuring Admission Control Using Cloudera Manager

In Cloudera Manager, you can configure pools to manage queued Impala queries, and the options for the limit on number of concurrent queries and how to handle queries that exceed the limit. For details, see [Managing Resources with Cloudera Manager](#).



See [Examples of Admission Control Configurations](#) on page 193 for a sample setup for admission control under Cloudera Manager.

### Configuring Admission Control Using the Command Line

If you do not use Cloudera Manager, you use a combination of startup options for the Impala daemon, and optionally editing or manually constructing the configuration files `fair-scheduler.xml` and `llama-site.xml`.

For a straightforward configuration using a single resource pool named `default`, you can specify configuration options on the command line and skip the `fair-scheduler.xml` and `llama-site.xml` configuration files.

For an advanced configuration with multiple resource pools using different settings, set up the `fair-scheduler.xml` and `llama-site.xml` configuration files manually. Provide the paths to each one using the Impala daemon command-line options, `--fair_scheduler_allocation_path` and `--llama_site_path` respectively.

The Impala admission control feature only uses the Fair Scheduler configuration settings to determine how to map users and groups to different resource pools. For example, you might set up different resource pools with separate memory limits, and maximum number of concurrent and queued queries, for different categories of users within your organization. For details about all the Fair Scheduler configuration settings, see the [Apache wiki](#).

The Impala admission control feature only uses a small subset of possible settings from the `llama-site.xml` configuration file:

```
llama.am.throttling.maximum.placed.reservations.queue_name
llama.am.throttling.maximum.queued.reservations.queue_name
```

For details about all the Llama configuration settings, see [Llama Default Configuration](#).

See [Example Admission Control Configurations Using Configuration Files](#) on page 195 for sample configuration files for admission control using multiple resource pools, without Cloudera Manager.

### Examples of Admission Control Configurations

#### Example Admission Control Configurations Using Cloudera Manager

For full instructions about configuring dynamic resource pools through Cloudera Manager, see [Dynamic Resource Pools](#) in the CDH 5 documentation. The following examples demonstrate some important points related to the Impala admission control feature.

The following figure shows a sample of the Dynamic Resource Pools page in Cloudera Manager, accessed through the **Clusters > Cluster name > Resource Management > Dynamic Resource Pools** menu choice and then the **Configuration** tab. Numbers from all the resource pools are combined into the topmost `root` pool. The `default` pool is for users who are not assigned any other pool by the user-to-pool mapping settings. The `development` and `production` pools show how you can set different limits for different classes of users, for total memory, number of concurrent queries, and number of queries that can be queued.

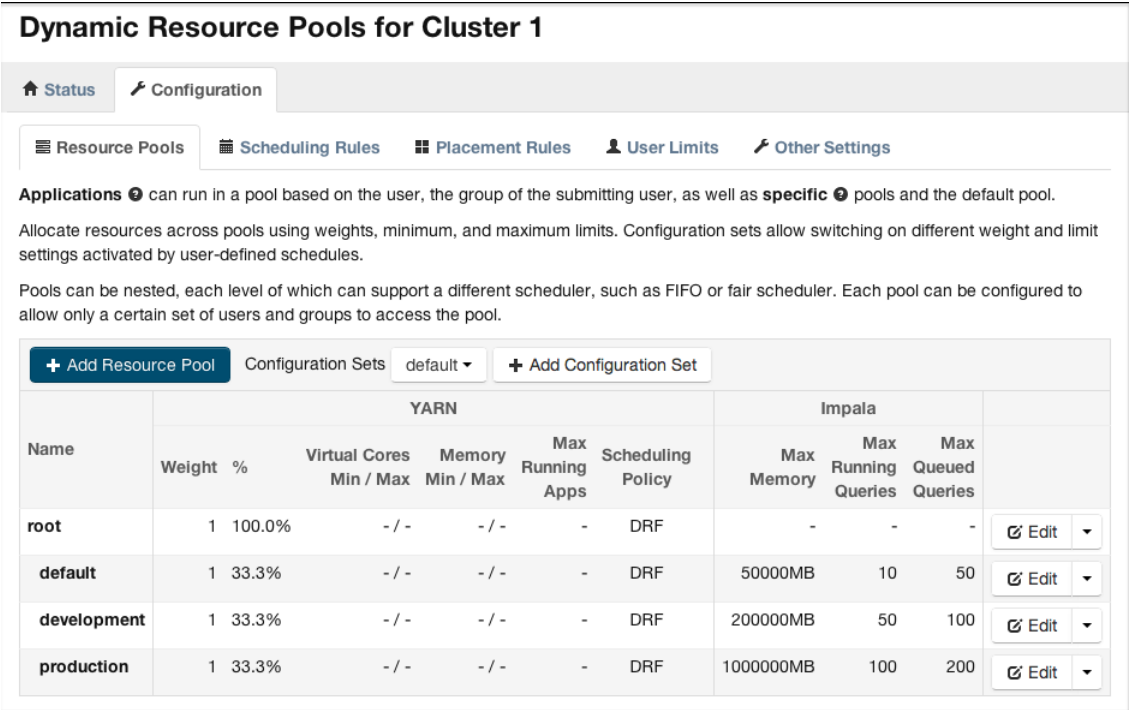
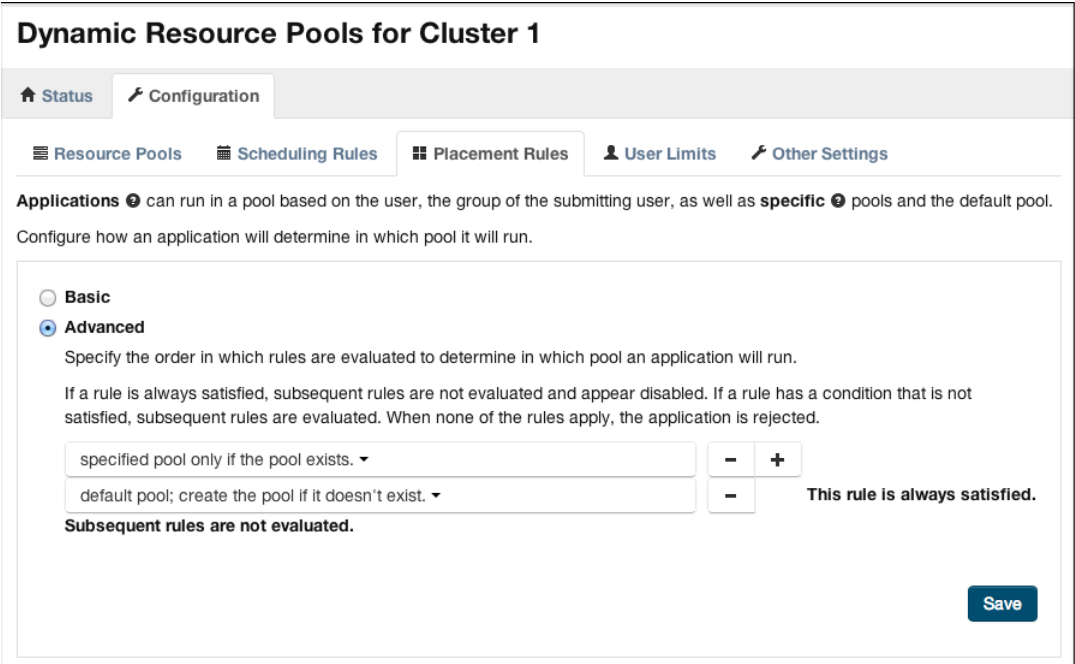


Figure 1: Sample Settings for Cloudera Manager Dynamic Resource Pools Page

The following figure shows a sample of the Placement Rules page in Cloudera Manager, accessed through the **Clusters > Cluster name > Resource Management > Dynamic Resource Pools** menu choice and then the **Configuration > Placement Rules** tabs. The settings demonstrate a reasonable configuration of a pool named `default` to service all requests where the specified resource pool does not exist, is not explicitly set, or the user or group is not authorized for the specified pool.



### Example Admission Control Configurations Using Configuration Files

For clusters not managed by Cloudera Manager, here are sample `fair-scheduler.xml` and `llama-site.xml` files that define resource pools equivalent to the ones in the preceding Cloudera Manager dialog. These sample files are stripped down: in a real deployment they might contain other settings for use with various aspects of the YARN and Llama components. The settings shown here are the significant ones for the Impala admission control feature.

#### **fair-scheduler.xml:**

Although Impala does not use the `vcores` value, you must still specify it to satisfy YARN requirements for the file contents.

Each `<aclSubmitApps>` tag (other than the one for `root`) contains a comma-separated list of users, then a space, then a comma-separated list of groups; these are the users and groups allowed to submit Impala statements to the corresponding resource pool.

If you leave the `<aclSubmitApps>` element empty for a pool, nobody can submit directly to that pool; child pools can specify their own `<aclSubmitApps>` values to authorize users and groups to submit to those pools.

```
<allocations>
  <queue name="root">
    <aclSubmitApps> </aclSubmitApps>
    <queue name="default">
      <maxResources>50000 mb, 0 vcores</maxResources>
      <aclSubmitApps>*</aclSubmitApps>
    </queue>
    <queue name="development">
      <maxResources>200000 mb, 0 vcores</maxResources>
      <aclSubmitApps>user1,user2 dev,ops,admin</aclSubmitApps>
    </queue>
    <queue name="production">
      <maxResources>1000000 mb, 0 vcores</maxResources>
      <aclSubmitApps> ops,admin</aclSubmitApps>
    </queue>
  </queue>
  <queuePlacementPolicy>
    <rule name="specified" create="false"/>
    <rule name="default" />
  </queuePlacementPolicy>
</allocations>
```

#### **llama-site.xml:**

```
<?xml version="1.0" encoding="UTF-8"?>
<configuration>
  <property>
    <name>llama.am.throttling.maximum.placed.reservations.root.default</name>
    <value>10</value>
  </property>
  <property>
    <name>llama.am.throttling.maximum.queued.reservations.root.default</name>
    <value>50</value>
  </property>
  <property>
    <name>llama.am.throttling.maximum.placed.reservations.root.development</name>
    <value>50</value>
  </property>
  <property>
    <name>llama.am.throttling.maximum.queued.reservations.root.development</name>
    <value>100</value>
  </property>
  <property>
    <name>llama.am.throttling.maximum.placed.reservations.root.production</name>
    <value>100</value>
  </property>
  <property>
    <name>llama.am.throttling.maximum.queued.reservations.root.production</name>
    <value>200</value>
  </property>
</configuration>
```

```
</property>  
</configuration>
```

### Guidelines for Using Admission Control

To see how admission control works for particular queries, examine the profile output for the query. This information is available through the `PROFILE` statement in `impala-shell` immediately after running a query in the shell, on the **queries** page of the Impala debug web UI, or in the Impala log file (basic information at log level 1, more detailed information at log level 2). The profile output contains details about the admission decision, such as whether the query was queued or not and which resource pool it was assigned to. It also includes the estimated and actual memory usage for the query, so you can fine-tune the configuration for the memory limits of the resource pools.

Where practical, use Cloudera Manager to configure the admission control parameters. The Cloudera Manager GUI is much simpler than editing the configuration files directly.

Remember that the limits imposed by admission control are “soft” limits. Although the limits you specify for number of concurrent queries and amount of memory apply cluster-wide, the decentralized nature of this mechanism means that each Impala node makes its own decisions about whether to allow queries to run immediately or to queue them. These decisions rely on information passed back and forth between nodes by the statestore service. If a sudden surge in requests causes more queries than anticipated to run concurrently, then as a fallback, the overall Impala memory limit and the Linux cgroups mechanism serve as hard limits to prevent overallocation of memory, by cancelling queries if necessary.

If you have trouble getting a query to run because its estimated memory usage is too high, you can override the estimate by setting the `MEM_LIMIT` query option in `impala-shell`, then issuing the query through the shell in the same session. The `MEM_LIMIT` value is treated as the estimated amount of memory, overriding the estimate that Impala would generate based on table and column statistics. This value is used only for making admission control decisions, and is not pre-allocated by the query.

In `impala-shell`, you can also specify which resource pool to direct queries to by setting the `REQUEST_POOL` query option. (This option was named `YARN_POOL` during the CDH 5 beta period.)

The statements affected by the admission control feature are primarily queries, but also include statements that write data such as `INSERT` and `CREATE TABLE AS SELECT`. Most write operations in Impala are not resource-intensive, but inserting into a Parquet table can require substantial memory due to buffering 1 GB of data before writing out each Parquet data block. See [Loading Data into Parquet Tables](#) for instructions about inserting data efficiently into Parquet tables.

Although admission control does not scrutinize memory usage for other kinds of DDL statements, if a query is queued due to a limit on concurrent queries or memory usage, subsequent statements in the same session are also queued so that they are processed in the correct order:

```
-- This query could be queued to avoid out-of-memory at times of heavy load.  
select * from huge_table join enormous_table using (id);  
-- If so, this subsequent statement in the same session is also queued  
-- until the previous statement completes.  
drop table huge_table;
```

If you set up different resource pools for different users and groups, consider reusing any classifications and hierarchy you developed for use with Sentry security. See [Enabling Sentry Authorization for Impala](#) for details.

For details about all the Fair Scheduler configuration settings, see [Fair Scheduler Configuration](#), in particular the tags such as `<queue>` and `<aclSubmitApps>` to map users and groups to particular resource pools (queues).

### Integrated Resource Management with YARN

You can limit the CPU and memory resources used by Impala, to manage and prioritize workloads on clusters that run jobs from many Hadoop components. (Currently, there is no limit or throttling on the I/O for Impala queries.) In CDH 5, Impala can use the underlying Apache Hadoop YARN resource management framework, which

allocates the required resources for each Impala query. Impala estimates the resources required by the query on each host of the cluster, and requests the resources from YARN.

Requests from Impala to YARN go through an intermediary service called Llama. When the resource requests are granted, Impala starts the query and places all relevant execution threads into the cgroup containers and sets up the memory limit on each host. If sufficient resources are not available, the Impala query waits until other jobs complete and the resources are freed. During query processing, as the need for additional resources arises, Llama can “expand” already-requested resources, to avoid over-allocating at the start of the query.

After a query is finished, Llama caches the resources (for example, leaving memory allocated) in case they are needed for subsequent Impala queries. This caching mechanism avoids the latency involved in making a whole new set of resource requests for each query. If the resources are needed by YARN for other types of jobs, Llama returns them.

While the delays to wait for resources might make individual queries seem less responsive on a heavily loaded cluster, the resource management feature makes the overall performance of the cluster smoother and more predictable, without sudden spikes in utilization due to memory paging, CPUs pegged at 100%, and so on.

## The Llama Daemon

Llama is a system that mediates resource management between Cloudera Impala and Hadoop YARN. Llama enables Impala to reserve, use, and release resource allocations in a Hadoop cluster. Llama is only required if resource management is enabled in Impala.

By default, YARN allocates resources bit-by-bit as needed by MapReduce jobs. Impala needs all resources available at the same time, so that intermediate results can be exchanged between cluster nodes, and queries do not stall partway through waiting for new resources to be allocated. Llama is the intermediary process that ensures all requested resources are available before each Impala query actually begins.

For management through Cloudera Manager, see [The Impala Llama ApplicationMaster](#).

## Controlling Resource Estimation Behavior

By default, Impala consults the table statistics and column statistics for each table in a query, and uses those figures to construct estimates of needed resources for each query. See [COMPUTE STATS Statement](#) for the statement to collect table and column statistics for a table.

To avoid problems with inaccurate or missing statistics, which can lead to inaccurate estimates of resource consumption, Impala allows you to set default estimates for CPU and memory resource consumption. As a query grows to require more resources, Impala will request more resources from Llama (this is called “expanding” a query reservation). When the query is complete, those resources are returned to YARN as normal. To enable this feature, use the command-line option `-rm_always_use_defaults` when starting `impalad`, and optionally `-rm_default_memory=size` and `-rm_default_cpu_cores`. Cloudera recommends always running with `-rm_always_use_defaults` enabled when using resource management, because if the query needs more resources than the default values, the resource requests are expanded dynamically as the query runs. See [impalad Startup Options for Resource Management](#) on page 198 for details about each option.

## Checking Resource Estimates and Actual Usage

To make resource usage easier to verify, the output of the `EXPLAIN SQL` statement now includes information about estimated memory usage, whether table and column statistics are available for each table, and the number of virtual cores that a query will use. You can get this information through the `EXPLAIN` statement without actually running the query. The extra information requires setting the query option `EXPLAIN_LEVEL=verbose`; see [EXPLAIN Statement](#) for details. The same extended information is shown at the start of the output from the `PROFILE` statement in `impala-shell`. The detailed profile information is only available after running the query. You can take appropriate actions (gathering statistics, adjusting query options) if you find that queries fail or run with suboptimal performance when resource management is enabled.

### How Resource Limits Are Enforced

- CPU limits are enforced by the Linux cgroups mechanism. YARN grants resources in the form of containers that correspond to cgroups on the respective machines.
- Memory is enforced by Impala's query memory limits. Once a reservation request has been granted, Impala sets the query memory limit according to the granted amount of memory before executing the query.

### Enabling Resource Management for Impala

To enable resource management for Impala, first you [set up the YARN and Llama services for your CDH cluster](#). Then you [add startup options and customize resource management settings](#) for the Impala services.

#### Required CDH Setup for Resource Management with Impala

YARN is the general-purpose service that manages resources for many Hadoop components within a CDH cluster. Llama is a specialized service that acts as an intermediary between Impala and YARN, translating Impala resource requests to YARN and coordinating with Impala so that queries only begin executing when all needed resources have been granted by YARN.

For information about setting up the YARN and Llama services, see the instructions for [Cloudera Manager](#).

#### Using Impala with a Llama High Availability Configuration

Impala can take advantage of the Llama high availability (HA) feature, with additional Llama servers that step in if the primary one becomes unavailable. (Only one Llama server at a time services all resource requests.) Before using this feature from Impala, read the background information about Llama HA, its main features, and how to set it up.

#### Command-line method for systems without Cloudera Manager:

Setting up the Impala side in a Llama HA configuration involves setting the `impalad` configuration options `-llama_addresses` (mandatory) and optionally `-llama_max_request_attempts`, `-llama_registration_timeout_secs`, and `-llama_registration_wait_secs`. See the next section [impalad Startup Options for Resource Management](#) on page 198 for usage instructions for those options.

The `impalad` daemon on the coordinator host registers with the Llama server for each query, receiving a handle that is used for subsequent resource requests. If a Llama server becomes unavailable, all running Impala queries are cancelled. Subsequent queries register with the next specified Llama server. This registration only happens when a query or similar request causes an `impalad` to request resources through Llama. Therefore, when a Llama server becomes unavailable, that fact might not be reported immediately in the Impala status information such as the `metrics` page in the debug web UI.

**Cloudera Manager method:** See [Llama High Availability](#).

#### impalad Startup Options for Resource Management

The following startup options for `impalad` enable resource management and customize its parameters for your cluster configuration:

- `-enable_rm`: Whether to enable resource management or not, either `true` or `false`. The default is `false`. None of the other resource management options have any effect unless `-enable_rm` is turned on.
- `-llama_host`: Hostname or IP address of the Llama service that Impala should connect to. The default is `127.0.0.1`.
- `-llama_port`: Port of the Llama service that Impala should connect to. The default is `15000`.
- `-llama_callback_port`: Port that Impala should start its Llama callback service on. Llama reports when resources are granted or preempted through that service.
- `-cgroup_hierarchy_path`: Path where YARN and Llama will create cgroups for granted resources. Impala assumes that the cgroup for an allocated container is created in the path `'cgroup_hierarchy_path + container_id'`.
- `-rm_always_use_defaults`: If this Boolean option is enabled, Impala ignores computed estimates and always obtains the default memory and CPU allocation from Llama at the start of the query. These default estimates are approximately 2 CPUs and 4 GB of memory, possibly varying slightly depending on cluster size,

workload, and so on. Cloudera recommends enabling `-rm_always_use_defaults` whenever resource management is used, and relying on these default values (that is, leaving out the two following options).

- `-rm_default_memory=size`: Optionally sets the default estimate for memory usage for each query. You can use suffixes such as M and G for megabytes and gigabytes, the same as with the [MEM\\_LIMIT](#) query option. Only has an effect when `-rm_always_use_defaults` is also enabled.
- `-rm_default_cpu_cores`: Optionally sets the default estimate for number of virtual CPU cores for each query. Only has an effect when `-rm_always_use_defaults` is also enabled.

The following options fine-tune the interaction of Impala with Llama when Llama high availability (HA) is enabled. The `-llama_addresses` option is only applicable in a Llama HA environment. `-llama_max_request_attempts`, `-llama_registration_timeout_secs`, and `-llama_registration_wait_secs` work whether or not Llama HA is enabled, but are most useful in combination when Llama is set up for high availability.

- `-llama_addresses`: Comma-separated list of `hostname:port` items, specifying all the members of the Llama availability group. Defaults to "127.0.0.1:15000".
- `-llama_max_request_attempts`: Maximum number of times a request to reserve, expand, or release resources is retried until the request is cancelled. Attempts are only counted after Impala is registered with Llama. That is, a request survives at most `llama_max_request_attempts-1` re-registrations. Defaults to 5.
- `-llama_registration_timeout_secs`: Maximum number of seconds that Impala will attempt to register or re-register with Llama. If registration is unsuccessful, Impala cancels the action with an error, which could result in an `impalad` startup failure or a cancelled query. A setting of -1 means try indefinitely. Defaults to 30.
- `-llama_registration_wait_secs`: Number of seconds to wait between attempts during Llama registration. Defaults to 3.

### impala-shell Query Options for Resource Management

Before issuing SQL statements through the `impala-shell` interpreter, you can use the `SET` command to configure the following parameters related to resource management:

- [EXPLAIN\\_LEVEL Query Option](#)
- [MEM\\_LIMIT Query Option](#)
- [RESERVATION\\_REQUEST\\_TIMEOUT Query Option \(CDH 5 only\)](#)
- [V\\_CPU\\_CORES Query Option \(CDH 5 only\)](#)

### Limitations of Resource Management for Impala

Currently, Impala in CDH 5 has the following limitations for resource management of Impala queries:

- Table statistics are required, and column statistics are highly valuable, for Impala to produce accurate estimates of how much memory to request from YARN. See [Overview of Table Statistics](#) and [Overview of Column Statistics](#) for instructions on gathering both kinds of statistics, and [EXPLAIN Statement](#) for the extended `EXPLAIN` output where you can check that statistics are available for a specific table and set of columns.
- If the Impala estimate of required memory is lower than is actually required for a query, Impala dynamically expands the amount of requested memory. Queries might still be cancelled if the reservation expansion fails, for example if there are insufficient remaining resources for that pool, or the expansion request takes long enough that it exceeds the query timeout interval, or because of YARN preemption. You can see the actual memory usage after a failed query by issuing a `PROFILE` command in `impala-shell`. Specify a larger memory figure with the `MEM_LIMIT` query option and re-try the query.

The `MEM_LIMIT` query option, and the other resource-related query options, are settable through the ODBC or JDBC interfaces in Impala 2.0 and higher. This is a former limitation that is now lifted.



# Performance Management

This section describes mechanisms and best practices for improving performance.

## Related Information

- [Tuning Impala for Performance](#)

## Improving Performance

This section provides solutions to some performance problems, and describes configuration best practices.

- **Important:** If you are running CDH over 10 Gbps Ethernet, improperly set network configuration or improperly applied NIC firmware or drivers can noticeably degrade performance. Work with your network engineers and hardware vendors to make sure that you have the proper NIC firmware, drivers, and configurations in place and that your network performs properly. Cloudera recognizes that network setup and upgrade are challenging problems, and will make best efforts to share any helpful experiences.

### Disabling Transparent Hugepage Compaction

Most Linux platforms supported by CDH 5 include a feature called **transparent hugepage compaction** which interacts poorly with Hadoop workloads and can seriously degrade performance.

**Symptom:** `top` and other system monitoring tools show a large percentage of the CPU usage classified as "system CPU". If system CPU usage is 30% or more of the total CPU usage, your system may be experiencing this issue.

#### What to do:

- **Note:** In the following instructions, `defrag_file_pathname` depends on your operating system:
  - Red Hat/CentOS: `/sys/kernel/mm/redhat_transparent_hugepage/defrag`
  - Ubuntu/Debian, OEL, SLES: `/sys/kernel/mm/transparent_hugepage/defrag`

1. To see whether transparent hugepage compaction is enabled, run the following command and check the output:

```
$ cat defrag_file_pathname
```

- `[always] never` means that transparent hugepage compaction is enabled.
- `always [never]` means that transparent hugepage compaction is disabled.

2. To disable transparent hugepage compaction, add the following command to `/etc/rc.local`:

```
echo never > defrag_file_pathname
```

You can also disable transparent hugepage compaction interactively (but remember this will not survive a reboot).

#### To disable transparent hugepage compaction temporarily as root:

```
# echo 'never' > defrag_file_pathname
```



**To disable transparent hugepage compaction temporarily using sudo:**

```
$ sudo sh -c "echo 'never' > defrag_file_pathname"
```

**Setting the vm.swappiness Linux Kernel Parameter**

`vm.swappiness` is a Linux kernel parameter that controls how aggressively memory pages are swapped to disk. It can be set to a value between 0-100; the higher the value, the more aggressive the kernel is in seeking out inactive memory pages and swapping them to disk.

You can see what value `vm.swappiness` is currently set to by looking at `/proc/sys/vm`; for example:

```
cat /proc/sys/vm/swappiness
```

On most systems, it is set to 60 by default. This is not suitable for Hadoop cluster nodes, because it can cause processes to get swapped out even when there is free memory available. This can affect stability and performance, and may cause problems such as lengthy garbage collection pauses for important system daemons. Cloudera recommends that you set this parameter to 10 or less; for example:

```
# sysctl -w vm.swappiness=10
```

Cloudera previously recommended a setting of 0, but in recent kernels (such as those included with RedHat 6.4 and higher, and Ubuntu 12.04 LTS and higher) a setting of 0 might lead to out of memory issues per this blog post: <http://www.percona.com/blog/2014/04/28/oom-relation-vm-swappiness0-new-kernel/>.

**Improving Performance in Shuffle Handler and IFile Reader**

The MapReduce shuffle handler and IFile reader use native Linux calls (`posix_fadvise(2)` and `sync_data_range`) on Linux systems with Hadoop native libraries installed. The subsections that follow provide details.

**Shuffle Handler**

You can improve MapReduce shuffle handler performance by enabling shuffle readahead. This causes the TaskTracker or Node Manager to pre-fetch map output before sending it over the socket to the reducer.

- To enable this feature for YARN, set the `mapreduce.shuffle.manage.os.cache` property to `true` (default). To further tune performance, adjust the value of the `mapreduce.shuffle.readahead.bytes` property. The default value is 4MB.
- To enable this feature for MRv1, set the `mapred.tasktracker.shuffle.fadvise` property to `true` (default). To further tune performance, adjust the value of the `mapred.tasktracker.shuffle.readahead.bytes` property. The default value is 4MB.

**IFile Reader**

Enabling IFile readahead increases the performance of merge operations. To enable this feature for either MRv1 or YARN, set the `mapreduce.ifile.readahead` property to `true` (default). To further tune the performance, adjust the value of the `mapreduce.ifile.readahead.bytes` property. The default value is 4MB.

**Best Practices for MapReduce Configuration**

The configuration settings described below can reduce inherent latencies in MapReduce execution. You set these values in `mapred-site.xml`.

**Send a heartbeat as soon as a task finishes**

Set the `mapreduce.tasktracker.outofband.heartbeat` property to `true` to let the TaskTracker send an out-of-band heartbeat on task completion to reduce latency; the default value is `false`:

```
<property>
  <name>mapreduce.tasktracker.outofband.heartbeat</name>
```

```
<value>true</value>
</property>
```

### Reduce the interval for JobClient status reports on single node systems

The `jobclient.progress.monitor.poll.interval` property defines the interval (in milliseconds) at which JobClient reports status to the console and checks for job completion. The default value is 1000 milliseconds; you may want to set this to a lower value to make tests run faster on a single-node cluster. Adjusting this value on a large production cluster may lead to unwanted client-server traffic.

```
<property>
  <name>jobclient.progress.monitor.poll.interval</name>
  <value>10</value>
</property>
```

### Tune the JobTracker heartbeat interval

Tuning the minimum interval for the TaskTracker-to-JobTracker heartbeat to a smaller value may improve MapReduce performance on small clusters.

```
<property>
  <name>mapreduce.jobtracker.heartbeat.interval.min</name>
  <value>10</value>
</property>
```

### Start MapReduce JVMs immediately

The `mapred.reduce.slowstart.completed.maps` property specifies the proportion of Map tasks in a job that must be completed before any Reduce tasks are scheduled. For small jobs that require fast turnaround, setting this value to 0 can improve performance; larger values (as high as 50%) may be appropriate for larger jobs.

```
<property>
  <name>mapred.reduce.slowstart.completed.maps</name>
  <value>0</value>
</property>
```

## Tips and Best Practices for Jobs

This section describes changes you can make at the job level.

### Use the Distributed Cache to Transfer the Job JAR

Use the distributed cache to transfer the job JAR rather than using the `JobConf(Class)` constructor and the `JobConf.setJar()` and `JobConf.setJarByClass()` methods.

To add JARs to the classpath, use `-libjars jar1,jar2`, which will copy the local JAR files to HDFS and then use the distributed cache mechanism to make sure they are available on the task nodes and are added to the task classpath.

The advantage of this over `JobConf.setJar` is that if the JAR is on a task node it won't need to be copied again if a second task from the same job runs on that node, though it will still need to be copied from the launch machine to HDFS.

- **Note:** `-libjars` works only if your MapReduce driver uses [ToolRunner](#). If it doesn't, you would need to use the DistributedCache APIs (Cloudera does not recommend this).

For more information, see item 1 in the blog post [How to Include Third-Party Libraries in Your MapReduce Job](#).

### Changing the Logging Level on a Job (MRv1)

You can change the logging level for an individual job. You do this by setting the following properties in the job configuration (`JobConf`):

- `mapreduce.map.log.level`
- `mapreduce.reduce.log.level`

Valid values are NONE, INFO, WARN, DEBUG, TRACE, and ALL.

**Example:**

```
JobConf conf = new JobConf();
...

conf.set("mapreduce.map.log.level", "DEBUG");
conf.set("mapreduce.reduce.log.level", "TRACE");
...
```

## Configuring Short-Circuit Reads

So-called "short-circuit" reads bypass the DataNode, allowing a client to read the file directly, as long as the client is co-located with the data. Short-circuit reads provide a substantial performance boost to many applications and help improve HBase random read profile and Impala performance.

Short-circuit reads require `libhadoop.so` (the [Hadoop Native Library](#)) to be accessible to both the server and the client. `libhadoop.so` is not available if you have installed from a tarball. You must install from an `.rpm`, `.deb`, or `parcel` in order to use short-circuit local reads.

### Configuring Short-Circuit Reads Using Cloudera Manager

**Required Role:** Configurator Cluster Administrator Full Administrator

- **Note:** Short-circuit reads are enabled by default in Cloudera Manager.

1. Go to the HDFS service.
2. Click the **Configuration** tab.
3. Type "shortcircuit" into the Search field to display the **Enable HDFS Short Circuit Read** property, and verify that this feature is enabled (set to True).
4. Go to the HBase service.
5. Click the **Configuration** tab.
6. Search for "shortcircuit".
7. Verify that the **Enable HDFS Short Circuit Read** property is enabled.

### Configuring Short-Circuit Reads Using the Command Line

Configure the following properties in `hdfs-site.xml`:

```
<property>
  <name>dfs.client.read.shortcircuit</name>
  <value>true</value>
</property>

<property>
  <name>dfs.client.read.shortcircuit.streams.cache.size</name>
  <value>1000</value>
</property>

<property>
  <name>dfs.client.read.shortcircuit.streams.cache.expiry.ms</name>
  <value>10000</value>
</property>

<property>
```

```
<name>dfs.domain.socket.path</name>  
<value>/var/run/hadoop-hdfs/dn._PORT</value>  
</property>
```

- **Note:** The text `_PORT` appears just as shown; you do not need to substitute a number.

If `/var/run/hadoop-hdfs/` is group-writable, make sure its group is `root`.

## Choosing a Data Compression Format

Whether to compress your data and which compression formats to use can have a significant impact on performance. Two of the most important places to consider data compression are in terms of MapReduce jobs and data stored in HBase. For the most part, the principles are similar for each.

### General Guidelines

- You need to balance the processing capacity required to compress and uncompress the data, the disk IO required to read and write the data, and the network bandwidth required to send the data across the network. The correct balance of these factors depends upon the characteristics of your cluster and your data, as well as your usage patterns.
- Compression is not recommended if your data is already compressed (such as images in JPEG format). In fact, the resulting file can actually be larger than the original.
- GZIP compression uses more CPU resources than Snappy or LZO, but provides a higher compression ratio. GZip is often a good choice for *cold data*, which is accessed infrequently. Snappy or LZO are a better choice for *hot data*, which is accessed frequently.
- BZip2 can also produce more compression than GZip for some types of files, at the cost of some speed when compressing and decompressing. HBase does not support BZip2 compression.
- Snappy often performs better than LZO. It is worth running tests to see if you detect a significant difference.
- For MapReduce, if you need your compressed data to be splittable, BZip2, LZO, and Snappy formats are splittable, but GZip is not. Splittability is not relevant to HBase data.
- For MapReduce, you can compress either the intermediate data, the output, or both. Adjust the parameters you provide for the MapReduce job accordingly. The following examples compress both the intermediate data and the output. MR2 is shown first, followed by MR1.

#### – MR2

```
hadoop jar hadoop-examples-.jar sort "-Dmapreduce.compress.map.output=true"  
"-Dmapreduce.map.output.compression.codec=org.apache.hadoop.io.compress.GzipCodec"  
"-Dmapreduce.output.compress=true"  
"-Dmapreduce.output.compression.codec=org.apache.hadoop.io.compress.GzipCodec"  
-outKey  
org.apache.hadoop.io.Text -outValue org.apache.hadoop.io.Text input output
```

#### – MR1

```
hadoop jar hadoop-examples-.jar sort "-Dmapred.compress.map.output=true"  
"-Dmapred.map.output.compression.codec=org.apache.hadoop.io.compress.GzipCodec"  
"-Dmapred.output.compress=true"  
"-Dmapred.output.compression.codec=org.apache.hadoop.io.compress.GzipCodec"  
-outKey  
org.apache.hadoop.io.Text -outValue org.apache.hadoop.io.Text input output
```

## Configuring Data Compression Using Cloudera Manager

To configure support for LZO using Cloudera Manager, you must install the GPL Extras package, then configure services to use it. See [Installing GPL Extras](#) and [Configuring Services to Use the GPL Extras Parcel](#) on page 95.

## Configuring Data Compression Using the Command Line

### ■ Important:

- If you use Cloudera Manager, do not use these command-line instructions.
- This information applies specifically to CDH 5.3.x. If you use an earlier version of CDH, see the documentation for that version located at [Cloudera Documentation](#).

To configure support for LZO in CDH, see [Step 5: \(Optional\) Install LZO](#) and [Configuring LZO](#). Snappy support is included in CDH.

To use Snappy in a MapReduce job, see [Using Snappy for MapReduce Compression](#). Use the same method for LZO, with the codec `com.hadoop.compression.lzo.LzopCodec` instead.

## Further Reading

For more information about compression algorithms in Hadoop, see section 4.2 of *Hadoop: The Definitive Guide*, by Tom White.

# Tuning the Solr Server

Solr performance tuning is a complex task. The following sections provide more details.

## Tuning to Complete During Setup

Some tuning is best completed during the setup of your system or may require some re-indexing.

### Configuring Lucene Version Requirements

You can configure Solr to use a specific version of Lucene. This can help ensure that the Lucene version that Search uses includes the latest features and bug fixes. At the time that a version of Solr ships, Solr is typically configured to use the appropriate Lucene version, in which case there is no need to change this setting. If a subsequent Lucene update occurs, you can configure the Lucene version requirements by directly editing the `luceneMatchVersion` element in the `solrconfig.xml` file. Versions are typically of the form `x.y`, such as `4.4`. For example, to specify version 4.4, you would ensure the following setting exists in `solrconfig.xml`:

```
<luceneMatchVersion>4.4</luceneMatchVersion>
```

### Designing the Schema

When constructing a schema, use data types that most accurately describe the data that the fields will contain. For example:

- Use the `tdate` type for dates. Do this instead of representing dates as strings.
- Consider using the `text` type that applies to your language, instead of using `String`. For example, you might use `text_en`. Text types support returning results for subsets of an entry. For example, querying on "john" would find "John Smith", whereas with the string type, only exact matches are returned.
- For IDs, use the string type.

### General Tuning

The following tuning categories can be completed at any time. It is less important to implement these changes before beginning to use your system.

#### General Tips

- Enabling multi-threaded faceting can provide better performance for field faceting. When multi-threaded faceting is enabled, field faceting tasks are completed in a parallel with a thread working on every field faceting task simultaneously. Performance improvements do not occur in all cases, but improvements are likely when all of the following are true:
  - The system uses highly concurrent hardware.
  - Faceting operations apply to large data sets over multiple fields.
  - There is not an unusually high number of queries occurring simultaneously on the system. Systems that are lightly loaded or that are mainly engaged with ingestion and indexing may be helped by multi-threaded faceting; for example, a system ingesting articles and being queried by a researcher. Systems heavily loaded by user queries are less likely to be helped by multi-threaded faceting; for example, an e-commerce site with heavy user-traffic.

- **Note:** Multi-threaded faceting only applies to field faceting and not to query faceting.
  - Field faceting identifies the number of unique entries for a field. For example, multi-threaded faceting could be used to simultaneously facet for the number of unique entries for the fields, "color" and "size". In such a case, there would be two threads, and each thread would work on faceting one of the two fields.
  - Query faceting identifies the number of unique entries that match a query for a field. For example, query faceting could be used to find the number of unique entries in the "size" field are between 1 and 5. Multi-threaded faceting does not apply to these operations.

To enable multi-threaded faceting, add `facet-threads` to queries. For example, to use up to 1000 threads, you might use a query as follows:

```
http://localhost:8983/solr/collection1/select?q=*:*&facet=true&fl=id&facet.field=f0_ws&facet.threads=1000
```

If `facet-threads` is omitted or set to 0, faceting is single-threaded. If `facet-threads` is set to a negative value, such as -1, multi-threaded faceting will use as many threads as there are fields to facet up to the maximum number of threads possible on the system.

- If your environment does not require Near Real Time (NRT), turn off soft auto-commit in `solrconfig.xml`.
- In most cases, do not change the default batch size setting of 1000. If you are working with especially large documents, you may consider decreasing the batch size.
- To help identify any garbage collector (GC) issues, enable GC logging in production. The overhead is low and the JVM supports GC log rolling as of 1.6.0\_34.
  - The minimum recommended GC logging flags are: `-XX:+PrintGCTimeStamps -XX:+PrintGCDateStamps -XX:+PrintGCDetails`.
  - To rotate the GC logs: `-Xloggc: -XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles= -XX:GCLogFileSize=`.

#### Solr and HDFS - the Block Cache

Cloudera Search enables Solr to store indexes in an HDFS filesystem. To maintain performance, an HDFS block cache has been implemented using Least Recently Used (LRU) semantics. This enables Solr to cache HDFS index files on read and write, storing the portions of the file in JVM "direct memory" (meaning off heap) by default or optionally in the JVM heap. Direct memory is preferred as it is not affected by garbage collection.

Batch jobs typically do not make use of the cache, while Solr servers (when serving queries or indexing documents) should. When running indexing using MapReduce, the MR jobs themselves do not make use of the block cache. Block caching is turned off by default and should be left disabled.

Tuning of this cache is complex and best practices are continually being refined. In general, allocate a cache that is about 10-20% of the amount of memory available on the system. For example, when running HDFS and Solr on a host with 50 GB of memory, typically allocate 5-10 GB of memory using `solr.hdfs.blockcache.slab.count`. As index sizes grow you may need to tune this parameter to maintain optimal performance.

- **Note:** Block cache metrics are currently unavailable.

### Configuration

The following parameters control caching. They can be configured at the Solr process level by setting the respective system property or by editing the `solrconfig.xml` directly.

Parameter	Default	Description
<code>solr.hdfs.blockcache.enabled</code>	true	Enable the block cache.
<code>solr.hdfs.blockcache.read.enabled</code>	true	Enable the read cache.
<code>solr.hdfs.blockcache.write.enabled</code>	false	Enable the write cache.
<code>solr.hdfs.blockcache.direct.memory.allocation</code>	true	Enable direct memory allocation. If this is false, heap is used.
<code>solr.hdfs.blockcache.slab.count</code>	1	Number of memory slabs to allocate. Each slab is 128 MB in size.
<code>solr.hdfs.blockcache.global</code>	true	If enabled, a single HDFS block cache is used for all SolrCores on a host. If <code>blockcache.global</code> is disabled, each SolrCore on a host creates its own private HDFS block cache. Enabling this parameter simplifies managing HDFS block cache memory.

- **Note:**

Increasing the direct memory cache size may make it necessary to increase the maximum direct memory size allowed by the JVM. Add the following to `/etc/default/solr` to do so. You must also replace `MAXMEM` with a reasonable upper limit. A typical default JVM value for this is 64 MB. When using `MAXMEM`, you must specify a unit such as `g` for gigabytes or `m` for megabytes. If `MAXMEM` were set to 2, the following command would set `MaxDirectMemorySize` to 2 GB:

```
CATALINA_OPTS="-XX:MaxDirectMemorySize=MAXMEMg -XX:+UseLargePages"
```

Restart Solr servers after editing this parameter.

Solr HDFS optimizes caching when performing NRT indexing using Lucene's `NRTCachingDirectory`.

Lucene caches a newly created segment if both of the following conditions are true:

- The segment is the result of a flush or a merge and the estimated size of the merged segment is  $\leq$  `solr.hdfs.nrtcachingdirectory.maxmergesizemb`.
- The total cached bytes is  $\leq$  `solr.hdfs.nrtcachingdirectory.maxcachedmb`.

The following parameters control NRT caching behavior:

Parameter	Default	Description
<code>solr.hdfs.nrtcachingdirectory.enable</code>	true	Whether to enable the NRTCachingDirectory.
<code>solr.hdfs.nrtcachingdirectory.maxcachedmb</code>	192	Size of the cache in megabytes.
<code>solr.hdfs.nrtcachingdirectory.maxmergesizemb</code>	16	Maximum segment size to cache.

Here is an example of `solrconfig.xml` with defaults:

```
<directoryFactory name="DirectoryFactory">
  <bool name="solr.hdfs.blockcache.enabled">${solr.hdfs.blockcache.enabled:true}</bool>

  <int
name="solr.hdfs.blockcache.slab.count">${solr.hdfs.blockcache.slab.count:1}</int>
  <bool
name="solr.hdfs.blockcache.direct.memory.allocation">${solr.hdfs.blockcache.direct.memory.allocation:true}</bool>

  <int
name="solr.hdfs.blockcache.blocksperbank">${solr.hdfs.blockcache.blocksperbank:16384}</int>

  <bool
name="solr.hdfs.blockcache.read.enabled">${solr.hdfs.blockcache.read.enabled:true}</bool>

  <bool
name="solr.hdfs.blockcache.write.enabled">${solr.hdfs.blockcache.write.enabled:true}</bool>

  <bool
name="solr.hdfs.nrtcachingdirectory.enable">${solr.hdfs.nrtcachingdirectory.enable:true}</bool>

  <int
name="solr.hdfs.nrtcachingdirectory.maxmergesizemb">${solr.hdfs.nrtcachingdirectory.maxmergesizemb:16}</int>

  <int
name="solr.hdfs.nrtcachingdirectory.maxcachedmb">${solr.hdfs.nrtcachingdirectory.maxcachedmb:192}</int>
</directoryFactory>
```

The following example illustrates passing Java options by editing the `/etc/default/solr` configuration file:

```
CATALINA_OPTS="-Xmx10g -XX:MaxDirectMemorySize=20g -XX:+UseLargePages
-Dsolr.hdfs.blockcache.slab.count=100"
```

For better performance, Cloudera recommends setting the Linux swap space on all Solr server hosts to 10 or less as shown below:

```
# set swappiness
sudo sysctl vm.swappiness=10
sudo bash -c 'echo "vm.swappiness=10">> /etc/sysctl.conf'
# disable swap space until next reboot:
sudo /sbin/swapoff -a
```

Cloudera previously recommended a setting of 0, but in recent kernels (such as those included with RedHat 6.4 and higher, and Ubuntu 12.04 LTS and higher) a setting of 0 might lead to out of memory issues per this blog post: <http://www.percona.com/blog/2014/04/28/oom-relation-vm-swappiness0-new-kernel/>.

## Threads

Configure the Tomcat server to have more threads per Solr instance. Note that this is only effective if your hardware is sufficiently powerful to accommodate the increased threads. 10,000 threads is a reasonable number to try in many cases.

To change the maximum number of threads, add a `maxThreads` element to the Connector definition in the Tomcat server's `server.xml` configuration file. For example, if you installed Search for CDH 5 using parcels



installation, you would modify the Connector definition in the `<parcel path>/CDH/etc/solr/tomcat-conf.dist/conf/server.xml` file so this:

```
<Connector port="${solr.port}" protocol="HTTP/1.1"
  connectionTimeout="20000"
  redirectPort="8443" />
```

Becomes this:

```
<Connector port="${solr.port}" protocol="HTTP/1.1"
  maxThreads="10000"
  connectionTimeout="20000"
  redirectPort="8443" />
```

## Garbage Collection

Choose different garbage collection options for best performance in different environments. Some garbage collection options typically chosen include:

- **Concurrent low pause collector:** Use this collector in most cases. This collector attempts to minimize "Stop the World" events. Avoiding these events can reduce connection timeouts, such as with ZooKeeper, and may improve user experience. This collector is enabled using `-XX:+UseConcMarkSweepGC`.
- **Throughput collector:** Consider this collector if raw throughput is more important than user experience. This collector typically uses more "Stop the World" events so this may negatively affect user experience and connection timeouts. This collector is enabled using `-XX:+UseParallelGC`.

You can also affect garbage collection behavior by increasing the Eden space to accommodate new objects. With additional Eden space, garbage collection does not need to run as frequently on new objects.

## Replicas

If you have sufficient additional hardware, add more replicas for a linear boost of query throughput. Note that adding replicas may slow write performance on the first replica, but otherwise this should have minimal negative consequences.

## Shards

In some cases, oversharding can help improve performance including intake speed. If your environment includes massively parallel hardware and you want to use these available resources, consider oversharding. You might increase the number of replicas per host from 1 to 2 or 3. Making such changes creates complex interactions, so you should continue to monitor your system's performance to ensure that the benefits of oversharding do not outweigh the costs.

## Commits

Changing commit values may improve performance in some situation. These changes result in tradeoffs and may not be beneficial in all cases.

- For hard commit values, the default value of 60000 (60 seconds) is typically effective, though changing this value to 120 seconds may improve performance in some cases. Note that setting this value to higher values, such as 600 seconds may result in undesirable performance tradeoffs.
- Consider increasing the auto-commit value from 15000 (15 seconds) to 120000 (120 seconds).
- Enable soft commits and set the value to the largest value that meets your requirements. The default value of 1000 (1 second) is too aggressive for some environments.

## Other Resources

- General information on Solr caching is available on the [SolrCaching](#) page on the Solr Wiki.

## Performance Management

- Information on issues that influence performance is available on the [SolrPerformanceFactors](#) page on the Solr Wiki.
- [Resource Management](#) describes how to use Cloudera Manager to manage resources, for example with Linux cgroups.
- For information on improving querying performance, see [ImproveSearchingSpeed](#).
- For information on improving indexing performance, see [ImproveIndexingSpeed](#).

# High Availability

This guide is for Apache Hadoop system administrators who want to enable continuous availability by configuring clusters without single points of failure.

## HDFS High Availability

This section explains how to configure a highly-available NameNode. The following topics provide more information and instructions:

### Introduction to HDFS High Availability

This section provides an overview of the HDFS high availability (HA) feature and how to configure and manage an HA HDFS cluster.

This document assumes that the reader has a general understanding of components and node types in an HDFS cluster. For details, see the [Apache HDFS Architecture Guide](#).

### Background

In a standard configuration, the NameNode is a single point of failure (SPOF) in an HDFS cluster. Each cluster has a single NameNode, and if that machine or process became unavailable, the cluster as a whole is unavailable until the NameNode is either restarted or brought up on a new host. The Secondary NameNode does not provide failover capability.

The standard configuration reduces the total availability of an HDFS cluster in two major ways:

- In the case of an unplanned event such as a host crash, the cluster is unavailable until an operator restarts the NameNode.
- Planned maintenance events such as software or hardware upgrades on the NameNode machine result in periods of cluster downtime.

HDFS HA addresses the above problems by providing the option of running two NameNodes in the same cluster, in an Active/Passive configuration. These are referred to as the Active NameNode and the Standby NameNode. Unlike the Secondary NameNode, the Standby NameNode is hot standby, allowing a fast failover to a new NameNode in the case that a machine crashes, or a graceful administrator-initiated failover for the purpose of planned maintenance. You cannot have more than two NameNodes.

### Implementation

Cloudera Manager 5 and CDH 5 support [Quorum-based Storage](#) on page 212 as the only HA implementation. In contrast, CDH 4 supports Quorum-based Storage and [shared storage using NFS](#). For instructions on switching to Quorum-based storage, see [Converting From an NFS-mounted Shared Edits Directory to Quorum-based Storage](#) on page 236.

### ■ Important:

- If you have upgraded from Cloudera Manager 4, and have a CDH 4 cluster using HA with an NFS-mounted shared edits directory, *your HA configuration will continue to work*. However, you will see a validation warning recommending you switch to Quorum-based storage.
- If you are using NFS-mounted shared edits directories and you disable HA, *you will not be able to re-enable HA in Cloudera Manager 5 using NFS-mounted shared directories*. Instead, you should configure HA to use Quorum-based storage.
- If you have HA enabled using an NFS-mounted shared edits directory, *you will be blocked from upgrading CDH 4 to CDH 5*. You must disable HA in order to perform the upgrade. After the upgrade, you will not be able to use NFS-mounted shared edits directories for edit log storage.
- If you are using CDH 4.0.x: CDH 4.0 did not support Quorum-based storage. Therefore, if you were using a HA configuration and you disable it, you will not be able to enable it through Cloudera Manager 5, since Cloudera Manager 5 does not support NFS-mounted storage. It is recommended that you upgrade your CDH 4.0 deployment to a more recent version of CDH.

## Quorum-based Storage

**Quorum-based Storage** refers to the HA implementation that uses a Quorum Journal Manager (QJM).

In order for the Standby NameNode to keep its state synchronized with the Active NameNode in this implementation, both nodes communicate with a group of separate daemons called JournalNodes. When any namespace modification is performed by the Active NameNode, it durably logs a record of the modification to a majority of these JournalNodes. The Standby NameNode is capable of reading the edits from the JournalNodes, and is constantly watching them for changes to the edit log. As the Standby Node sees the edits, it applies them to its own namespace. In the event of a failover, the Standby will ensure that it has read all of the edits from the JournalNodes before promoting itself to the Active state. This ensures that the namespace state is fully synchronized before a failover occurs.

In order to provide a fast failover, it is also necessary that the Standby NameNode has up-to-date information regarding the location of blocks in the cluster. In order to achieve this, the DataNodes are configured with the location of both NameNodes, and they send block location information and heartbeats to both.

It is vital for the correct operation of an HA cluster that only one of the NameNodes be active at a time. Otherwise, the namespace state would quickly diverge between the two, risking data loss or other incorrect results. In order to ensure this property and prevent the so-called "split-brain scenario," the JournalNodes will only ever allow a single NameNode to be a writer at a time. During a failover, the NameNode which is to become active will simply take over the role of writing to the JournalNodes, which will effectively prevent the other NameNode from continuing in the Active state, allowing the new Active NameNode to safely proceed with failover.

- **Note:** Because of this, fencing is not required, but it is still useful; see [Enabling HDFS HA](#) on page 213.

## Shared Storage Using NFS

In order for the Standby node to keep its state synchronized with the Active node, this implementation requires that the two nodes both have access to a directory on a shared storage device (for example, an NFS mount from a NAS).

When any namespace modification is performed by the Active node, it durably logs a record of the modification to an edit log file stored in the shared directory. The Standby node constantly watches this directory for edits, and when edits occur, the Standby node applies them to its own namespace. In the event of a failover, the Standby will ensure that it has read all of the edits from the shared storage before promoting itself to the Active state. This ensures that the namespace state is fully synchronized before a failover occurs.

In order to provide a fast failover, it is also necessary that the Standby node has up-to-date information regarding the location of blocks in the cluster. In order to achieve this, the DataNodes are configured with the location of both NameNodes, and they send block location information and heartbeats to both.

It is vital for the correct operation of an HA cluster that only one of the NameNodes be active at a time. Otherwise, the namespace state would quickly diverge between the two, risking data loss or other incorrect results. In order to ensure this and prevent the so-called "split-brain scenario," an administrator must configure at least one fencing method for the shared storage. During a failover, if it cannot be verified that the previous Active NameNode has relinquished its Active state, the fencing process is responsible for cutting off the previous Active NameNode's access to the shared edits storage. This prevents it from making any further edits to the namespace, allowing the new Active NameNode to safely proceed with failover.

## Configuring Hardware for HDFS HA

In order to deploy an HA cluster using Quorum-based Storage, you should prepare the following:

- NameNode machines - the machines on which you run the Active and Standby NameNodes should have equivalent hardware to each other, and equivalent hardware to what would be used in a non-HA cluster.
- JournalNode machines - the machines on which you run the JournalNodes.
  - Cloudera recommends that you deploy the JournalNode daemons on the "master" host or hosts (NameNode, Standby NameNode, JobTracker, and so on) so the JournalNodes' local directories can use the reliable local storage on those machines. Each JournalNode process should preferably have its own dedicated disk. You should not use SAN or NAS storage for these directories.
- There must be at least three JournalNode daemons, since edit log modifications must be written to a majority of JournalNodes. This will allow the system to tolerate the failure of a single machine. You can also run more than three JournalNodes, but in order to actually increase the number of failures the system can tolerate, you should run an odd number of JournalNodes, (three, five, seven, and so on). Note that when running with N JournalNodes, the system can tolerate at most  $(N - 1) / 2$  failures and continue to function normally. If the requisite quorum is not available, the NameNode will not format or start, and you will see an error similar to this:

```
12/10/01 17:34:18 WARN namenode.FSEditLog: Unable to determine input streams from QJM
to [10.0.1.10:8485, 10.0.1.10:8486, 10.0.1.10:8487]. Skipping.
java.io.IOException: Timed out waiting 20000ms for a quorum of nodes to respond.
```

- **Note:** In an HA cluster, the Standby NameNode also performs checkpoints of the namespace state, and thus it is not necessary to run a Secondary NameNode, CheckpointNode, or BackupNode in an HA cluster. In fact, to do so would be an error. If you are reconfiguring a non-HA-enabled HDFS cluster to be HA-enabled, you can reuse the hardware which you had previously dedicated to the Secondary NameNode.

## Enabling HDFS HA

An HDFS high availability (HA) cluster uses two NameNodes—an active NameNode and a standby NameNode. Only one NameNode can be active at any point in time. HDFS HA depends on maintaining a log of all namespace modifications in a location available to both NameNodes, so that in the event of a failure the standby NameNode has up-to-date information about the edits and location of blocks in the cluster. For CDH 4 HA features, see the [CDH 4 High Availability Guide](#).

- **Important:** When you enable or disable HA, the HDFS service and the services that depend on it—MapReduce, YARN, and HBase—are shut down. Therefore, you should not do this while you have jobs running on your cluster.

## Enabling HDFS HA Using Cloudera Manager

**Required Role:** Cluster Administrator Full Administrator

You can use Cloudera Manager to configure your CDH 4 or CDH 5 cluster for HDFS HA and automatic failover. In Cloudera Manager 5, HA is implemented using Quorum-based storage. Quorum-based storage relies upon a

set of JournalNodes, each of which maintains a local edits directory that logs the modifications to the namespace metadata. Enabling HA enables automatic failover as part of the same command.

■ **Important:**

- Enabling or disabling HA causes the previous monitoring history to become unavailable.
- Some parameters will be automatically set as follows once you have [enabled JobTracker HA](#). If you want to change the value from the default for these parameters, use an advanced configuration snippet.
  - `mapred.jobtracker.restart.recover: true`
  - `mapred.job.tracker.persist.jobstatus.active: true`
  - `mapred.ha.automatic-failover.enabled: true`
  - `mapred.ha.fencing.methods: shell(/bin/true)`

### Enabling High Availability and Automatic Failover

Make sure you have performed all the configuration and setup tasks described under [Configuring Hardware for HDFS HA](#) on page 213.

The **Enable High Availability** workflow leads you through adding a second (standby) NameNode and configuring JournalNodes. During the workflow, Cloudera Manager creates a [federated namespace](#).

1. Go to the HDFS service.
2. Select **Actions > Enable High Availability**. A screen showing the hosts that are eligible to run a standby NameNode and the JournalNodes displays.
  - a. Specify a name for the nameservice or accept the default name **nameservice1** and click **Continue**.
  - b. In the **NameNode Hosts** field, click **Select a host**. The host selection dialog displays.
  - c. Check the checkbox next to the hosts where you want the standby NameNode to be set up and click **OK**. The standby NameNode cannot be on the same host as the active NameNode, and the host that is chosen should have the same hardware configuration (RAM, disk space, number of cores, and so on) as the active NameNode.
  - d. In the **JournalNode Hosts** field, click **Select hosts**. The host selection dialog displays.
  - e. Check the checkboxes next to an odd number of hosts (a minimum of three) to act as JournalNodes and click **OK**. JournalNodes should be hosted on hosts with similar hardware specification as the NameNodes. It is recommended that you put a JournalNode each on the same hosts as the active and standby NameNodes, and the third JournalNode on similar hardware, such as the JobTracker.
  - f. Click **Continue**.
  - g. In the **JournalNode Edits Directory** property, enter a directory location for the JournalNode edits directory into the fields for each JournalNode host.
    - You may enter only one directory for each JournalNode. The paths do not need to be the same on every JournalNode.
    - The directories you specify should be empty, and must have the appropriate permissions.
  - h. **Extra Options:** Decide whether Cloudera Manager should clear existing data in ZooKeeper, standby NameNode, and JournalNodes. If the directories are not empty (for example, you are re-enabling a previous HA configuration), Cloudera Manager will not automatically delete the contents—you can select to delete the contents by keeping the default checkbox selection. The recommended default is to clear the directories. If you choose not to do so, the data should be in sync across the edits directories of the JournalNodes and should have the same version data as the NameNodes.
  - i. Click **Continue**.

Cloudera Manager executes a set of commands that will stop the dependent services, delete, create, and configure roles and directories as appropriate, create a nameservice and failover controller, and restart the dependent services and deploy the new client configuration.

3. If you want to use Hive, Impala, or Hue in a cluster with HA configured, follow the procedures in [Configuring Other CDH Components to Use HDFS HA](#) on page 231.
4. If you are running CDH 4.0 or 4.1, the standby NameNode may fail at the `bootstrapStandby` command with the error `Unable to read transaction ids 1-7 from the configured shared edits storage`. Use `rsync` or a similar tool to copy the contents of the `dfs.name.dir` directory from the active NameNode to the standby NameNode and start the standby NameNode.

- **Important:** If you change the NameNode Service RPC Port (`dfs.namenode.servicerpc-address`) while automatic failover is enabled, this will cause a mismatch between the NameNode address saved in the ZooKeeper `/hadoop-ha` znode and the NameNode address that the Failover Controller is configured with. This will prevent the Failover Controllers from restarting. If you need to change the NameNode Service RPC Port after Auto Failover has been enabled, you must do the following to re-initialize the znode:
  1. Stop the HDFS service.
  2. Configure the service RPC port:
    - a. Go to the HDFS service.
    - b. Click the **Configuration** tab.
    - c. Search for `dfs.namenode.servicerpc` which should display the **NameNode Service RPC Port** property. (It is found under the **NameNode Default Group** role group, **Ports and Addresses** category).
    - d. Change the port value as needed.
  3. On a ZooKeeper server host, run the ZooKeeper client CLI:
    - **Parcels** - `/opt/cloudera/parcels/CDH/lib/zookeeper/bin/zkCli.sh`
    - **Packages** - `/usr/lib/zookeeper/bin/zkCli.sh`
  4. Execute the following to remove the pre-configured nameservice. This example assumes the name of the Nameservice is **nameservice1**. You can identify the nameservice from the **Federation and High Availability** section on the **HDFS Instances** tab:
 

```
rmr /hadoop-ha/nameservice1
```
  5. Click the **HDFS Instances** tab.
  6. Select **Actions > Initialize High Availability State in ZooKeeper**.
  7. Start the HDFS service.

### Manually Failing Over to the Standby NameNode

If you are running an HDFS service with HA enabled, you can manually cause the active NameNode to failover to the standby NameNode. This is useful for planned downtime—for hardware changes, configuration changes, or software upgrades of your primary host.

1. Go to the HDFS service.
2. Click the **Instances** tab.
3. Select **Actions > Manual Failover**. (This option does not appear if HA is not enabled for the cluster.)
4. From the pop-up, select the NameNode that should be made active, then click **Manual Failover**.

- **Note: For advanced use only:** To force the selected NameNode to be active, irrespective of its state or the other NameNode state, set the **Force Failover** checkbox. You should choose this option only if automatic failover is *not enabled*. Forcing a failover will first attempt to failover the selected NameNode to active mode and the other NameNode to standby mode. It will do so even if the selected NameNode is in safe mode. If this fails, it will proceed to transition the selected NameNode to active mode. To avoid having two NameNodes be active, use this only if the other NameNode is either definitely stopped, or can be transitioned to standby mode by the first failover step.

5. When all the steps have been completed, click **Finish**.

Cloudera Manager transitions the NameNode you selected to be the active NameNode, and the other NameNode to be the standby NameNode. HDFS should *never* have two active NameNodes.

### Fencing Methods

In order to ensure that only one NameNode is active at a time, a fencing method is required for the shared edits directory. During a failover, the fencing method is responsible for ensuring that the previous active NameNode no longer has access to the shared edits directory, so that the new active NameNode can safely proceed writing to it.

By default, Cloudera Manager configures HDFS to use a shell fencing method (`shell(/cloudera_manager_agent_fencer.py)`) that takes advantage of the Cloudera Manager Agent. However, you can configure HDFS to use the `sshfence` method, or you can add your own shell fencing scripts, instead of or in addition to the one Cloudera Manager provides.

The fencing parameters are found in the **Service-Wide > High Availability** category under the configuration properties for your HDFS service.

For details of the fencing methods supplied with CDH 5, and how fencing is configured, see [Fencing Configuration](#) on page 219.

### Enabling HDFS HA Using the Command Line

#### ■ Important:

- If you use Cloudera Manager, do not use these command-line instructions.
- This information applies specifically to CDH 5.3.x. If you use an earlier version of CDH, see the documentation for that version located at [Cloudera Documentation](#).

This section describes the software configuration required for HDFS HA in CDH 5 and explains how to set configuration properties and use the command line to deploy HDFS HA.

### Configuring Software for HDFS HA

#### Configuration Overview

As with HDFS Federation configuration, HA configuration is backward compatible and allows existing single NameNode configurations to work without change. The new configuration is designed such that all the nodes in the cluster can have the same configuration without the need for deploying different configuration files to different machines based on the type of the node.

HA clusters reuse the **Nameservice ID** to identify a single HDFS instance that may consist of multiple HA NameNodes. In addition, there is a new abstraction called **NameNode ID**. Each distinct NameNode in the cluster has a different NameNode ID. To support a single configuration file for all of the NameNodes, the relevant configuration parameters include the Nameservice ID as well as the NameNode ID.

#### Changes to Existing Configuration Parameters

The following configuration parameter has changed for YARN implementations:

`fs.defaultFS` - formerly `fs.default.name`, the default path prefix used by the Hadoop FS client when none is given. (`fs.default.name` is deprecated for YARN implementations, but will still work.)

Optionally, you can configure the default path for Hadoop clients to use the HA-enabled logical URI. For example, if you use `mycluster` as the Nameservice ID as shown below, this will be the value of the authority portion of all of your HDFS paths. You can configure the default path in your `core-site.xml` file:



- For YARN:

```
<property>
  <name>fs.defaultFS</name>
  <value>hdfs://mycluster</value>
</property>
```

- For MRv1:

```
<property>
  <name>fs.default.name</name>
  <value>hdfs://mycluster</value>
</property>
```

## New Configuration Parameters

To configure HA NameNodes, you must add several configuration options to your `hdfs-site.xml` configuration file.

The order in which you set these configurations is unimportant, but the values you choose for `dfs.nameservices` and `dfs.ha.namenodes.[Nameservice ID]` will determine the keys of those that follow. This means that you should decide on these values before setting the rest of the configuration options.

### Configure `dfs.nameservices`

`dfs.nameservices` - the logical name for this new nameservice

Choose a logical name for this nameservice, for example `mycluster`, and use this logical name for the value of this configuration option. The name you choose is arbitrary. It will be used both for configuration and as the authority component of absolute HDFS paths in the cluster.

- **Note:** If you are also using HDFS Federation, this configuration setting should also include the list of other Nameservices, HA or otherwise, as a comma-separated list.

```
<property>
  <name>dfs.nameservices</name>
  <value>mycluster</value>
</property>
```

### Configure `dfs.ha.namenodes.[nameservice ID]`

`dfs.ha.namenodes.[nameservice ID]` - unique identifiers for each NameNode in the nameservice

Configure a list of comma-separated NameNode IDs. This will be used by DataNodes to determine all the NameNodes in the cluster. For example, if you used `mycluster` as the NameService ID previously, and you wanted to use `nn1` and `nn2` as the individual IDs of the NameNodes, you would configure this as follows:

```
<property>
  <name>dfs.ha.namenodes.mycluster</name>
  <value>nn1,nn2</value>
</property>
```

- **Note:** In this release, you can configure a maximum of two NameNodes per nameservice.

### Configure `dfs.namenode.rpc-address.[nameservice ID]`

`dfs.namenode.rpc-address.[nameservice ID].[name node ID]` - the fully-qualified RPC address for each NameNode to listen on

For both of the previously-configured NameNode IDs, set the full address and RPC port of the NameNode process. Note that this results in two separate configuration options. For example:

```
<property>
  <name>dfs.namenode.rpc-address.mycluster.nn1</name>
  <value>machine1.example.com:8020</value>
</property>
<property>
  <name>dfs.namenode.rpc-address.mycluster.nn2</name>
  <value>machine2.example.com:8020</value>
</property>
```

- **Note:** If necessary, you can similarly configure the `servicerpc-address` setting.

### Configure `dfs.namenode.http-address.[nameservice ID]`

`dfs.namenode.http-address.[nameservice ID].[name node ID]` - the fully-qualified HTTP address for each NameNode to listen on

Similarly `torpc-address` above, set the addresses for both NameNodes' HTTP servers to listen on. For example:

```
<property>
  <name>dfs.namenode.http-address.mycluster.nn1</name>
  <value>machine1.example.com:50070</value>
</property>
<property>
  <name>dfs.namenode.http-address.mycluster.nn2</name>
  <value>machine2.example.com:50070</value>
</property>
```

- **Note:** If you have Hadoop Kerberos security features enabled, and you intend to use HSFTP, you should also set the `https-address` similarly for each NameNode.

### Configure `dfs.namenode.shared.edits.dir`

`dfs.namenode.shared.edits.dir` - the location of the shared storage directory

Configure the addresses of the JournalNodes which provide the shared edits storage, written to by the Active NameNode and read by the Standby NameNode to stay up-to-date with all the file system changes the Active NameNode makes. Though you must specify several JournalNode addresses, **you should only configure one of these URIs**. The URI should be in the form:

```
qjournal://<host1:port1>;<host2:port2>;<host3:port3>/<journalId>
```

The Journal ID is a unique identifier for this nameservice, which allows a single set of JournalNodes to provide storage for multiple federated namesystems. Though it is not a requirement, it's a good idea to reuse the Nameservice ID for the journal identifier.

For example, if the JournalNodes for this cluster were running on the machines `node1.example.com`, `node2.example.com`, and `node3.example.com`, and the nameservice ID were `mycluster`, you would use the following as the value for this setting (the default port for the JournalNode is 8485):

```
<property>
  <name>dfs.namenode.shared.edits.dir</name>
  <value>qjournal://node1.example.com:8485;node2.example.com:8485;node3.example.com:8485/mycluster</value>
</property>
```

### Configure `dfs.journalnode.edits.dir`

`dfs.journalnode.edits.dir` - the path where the JournalNode daemon will store its local state

On each JournalNode machine, configure the absolute path where the edits and other local state information used by the JournalNodes will be stored; use only a single path per JournalNode. (The other JournalNodes provide redundancy; you can also configure this directory on a locally-attached RAID-1 or RAID-10 array.)

For example:

```
<property>
  <name>dfs.journalnode.edits.dir</name>
  <value>/data/1/dfs/jn</value>
</property>
```

Now create the directory (if it doesn't already exist) and make sure its owner is `hdfs`, for example:

```
$ sudo mkdir -p /data/1/dfs/jn
$ sudo chown -R hdfs:hdfs /data/1/dfs/jn
```

### Client Failover Configuration

`dfs.client.failover.proxy.provider.[nameservice ID]` - the Java class that HDFS clients use to contact the Active NameNode

Configure the name of the Java class which the DFS client will use to determine which NameNode is the current active, and therefore which NameNode is currently serving client requests. The only implementation which currently ships with Hadoop is the `ConfiguredFailoverProxyProvider`, so use this unless you are using a custom one. For example:

```
<property>
  <name>dfs.client.failover.proxy.provider.mycluster</name>
  <value>org.apache.hadoop.hdfs.server.namenode.ha.ConfiguredFailoverProxyProvider</value>
</property>
```

### Fencing Configuration

`dfs.ha.fencing.methods` - a list of scripts or Java classes which will be used to fence the active NameNode during a failover

It is desirable for correctness of the system that only one NameNode be in the active state at any given time.

- **Important:** When you use Quorum-based Storage, only one NameNode will ever be allowed to write to the JournalNodes, so there is no potential for corrupting the file system metadata in a "split-brain" scenario. But when a failover occurs, it is still possible that the previously active NameNode could serve read requests to clients - and these requests may be out of date - until that NameNode shuts down when it tries to write to the JournalNodes. For this reason, it is still desirable to configure some fencing methods even when using Quorum-based Storage.

To improve the availability of the system in the event the fencing mechanisms fail, it is advisable to configure a fencing method which is guaranteed to return success as the last fencing method in the list.

- **Note:** If you choose to use no actual fencing methods, you still must configure something for this setting, for example `shell(/bin/true)`.

The fencing methods used during a failover are configured as a carriage-return-separated list, and these will be attempted in order until one of them indicates that fencing has succeeded.

There are two fencing methods which ship with Hadoop:

- [sshfence](#)
- [shell](#)

For information on implementing your own custom fencing method, see the `org.apache.hadoop.ha.NodeFencer` class.

### Configuring the sshfence fencing method

`sshfence` - SSH to the active NameNode and kill the process

The `sshfence` option uses SSH to connect to the target node and uses `fuser` to kill the process listening on the service's TCP port. In order for this fencing option to work, it must be able to SSH to the target node without providing a passphrase. Thus, you must also configure the `dfs.ha.fencing.ssh.private-key-files` option, which is a comma-separated list of SSH private key files.

- **Important:** The files must be accessible to the user running the NameNode processes (typically the `hdfs` user on the NameNode hosts).

For example:

```
<property>
  <name>dfs.ha.fencing.methods</name>
  <value>sshfence</value>
</property>

<property>
  <name>dfs.ha.fencing.ssh.private-key-files</name>
  <value>/home/exampleuser/.ssh/id_rsa</value>
</property>
```

Optionally, you can configure a non-standard username or port to perform the SSH as shown below. You can also configure a timeout, in milliseconds, for the SSH, after which this fencing method will be considered to have failed:

```
<property>
  <name>dfs.ha.fencing.methods</name>
  <value>sshfence([[username][:port]])</value>
</property>
<property>
  <name>dfs.ha.fencing.ssh.connect-timeout</name>
  <value>30000</value>
  <description>
    SSH connection timeout, in milliseconds, to use with the builtin
    sshfence fencer.
  </description>
</property>
```

### Configuring the shell fencing method

`shell` - run an arbitrary shell command to fence the active NameNode

The shell fencing method runs an arbitrary shell command, which you can configure as shown below:

```
<property>
  <name>dfs.ha.fencing.methods</name>
  <value>shell(/path/to/my/script.sh arg1 arg2 ...)</value>
</property>
```

The string between '(' and ')' is passed directly to a `bash` shell and cannot include any closing parentheses.

When executed, the first argument to the configured script will be the address of the NameNode to be fenced, followed by all arguments specified in the configuration.

The shell command will be run with an environment set up to contain all of the current Hadoop configuration variables, with the '\_' character replacing any '.' characters in the configuration keys. The configuration used has already had any NameNode-specific configurations promoted to their generic forms - for example `dfs_namenode_rpc-address` will contain the RPC address of the target node, even though the configuration may specify that variable as `dfs.namenode.rpc-address.ns1.nn1`.

The following variables referring to the target node to be fenced are also available:

Variable	Description
\$target_host	Hostname of the node to be fenced
\$target_port	IPC port of the node to be fenced
\$target_address	The above two variables, combined as <i>host:port</i>
\$target_nameserviceid	The nameservice ID of the NameNode to be fenced
\$target_namenodeid	The NameNode ID of the NameNode to be fenced

You can also use these environment variables as substitutions in the shell command itself. For example:

```
<property>
  <name>dfs.ha.fencing.methods</name>
  <value>shell(/path/to/my/script.sh --nameservice=$target_nameserviceid
    $target_host:$target_port)</value>
</property>
```

If the shell command returns an exit code of 0, the fencing is determined to be successful. If it returns any other exit code, the fencing was not successful and the next fencing method in the list will be attempted.

- **Note:** This fencing method does not implement any timeout. If timeouts are necessary, they should be implemented in the shell script itself (for example, by forking a subshell to kill its parent in some number of seconds).

## Automatic Failover Configuration

The above sections describe how to configure manual failover. In that mode, the system will not automatically trigger a failover from the active to the standby NameNode, even if the active node has failed. This section describes how to configure and deploy automatic failover.

### Component Overview

Automatic failover adds two new components to an HDFS deployment: a ZooKeeper quorum, and the `ZKFailoverController` process (abbreviated as ZKFC).

Apache ZooKeeper is a highly available service for maintaining small amounts of coordination data, notifying clients of changes in that data, and monitoring clients for failures. The implementation of automatic HDFS failover relies on ZooKeeper for the following things:

- **Failure detection** - each of the NameNode machines in the cluster maintains a persistent session in ZooKeeper. If the machine crashes, the ZooKeeper session will expire, notifying the other NameNode that a failover should be triggered.
- **Active NameNode election** - ZooKeeper provides a simple mechanism to exclusively elect a node as active. If the current active NameNode crashes, another node can take a special exclusive lock in ZooKeeper indicating that it should become the next active NameNode.

The `ZKFailoverController` (ZKFC) is a new component - a ZooKeeper client which also monitors and manages the state of the NameNode. Each of the machines which runs a NameNode also runs a ZKFC, and that ZKFC is responsible for:

- **Health monitoring** - the ZKFC pings its local NameNode on a periodic basis with a health-check command. So long as the NameNode responds promptly with a healthy status, the ZKFC considers the node healthy. If the node has crashed, frozen, or otherwise entered an unhealthy state, the health monitor will mark it as unhealthy.
- **ZooKeeper session management** - when the local NameNode is healthy, the ZKFC holds a session open in ZooKeeper. If the local NameNode is active, it also holds a special lock `znode`. This lock uses ZooKeeper's support for "ephemeral" nodes; if the session expires, the lock node will be automatically deleted.

- **ZooKeeper-based election** – if the local NameNode is healthy, and the ZKFC sees that no other node currently holds the lock znode, it will itself try to acquire the lock. If it succeeds, then it has "won the election", and is responsible for running a failover to make its local NameNode active. The failover process is similar to the manual failover described above: first, the previous active is fenced if necessary, and then the local NameNode transitions to active state.

### Deploying ZooKeeper

In a typical deployment, ZooKeeper daemons are configured to run on three or five nodes. Since ZooKeeper itself has light resource requirements, it is acceptable to colocate the ZooKeeper nodes on the same hardware as the HDFS NameNode and Standby Node. Operators using MapReduce v2 (MRv2) often choose to deploy the third ZooKeeper process on the same node as the YARN ResourceManager. It is advisable to configure the ZooKeeper nodes to store their data on separate disk drives from the HDFS metadata for best performance and isolation.

See the [ZooKeeper documentation](#) for instructions on how to set up a ZooKeeper ensemble. In the following sections we assume that you have set up a ZooKeeper cluster running on three or more nodes, and have verified its correct operation by connecting using the ZooKeeper command-line interface (CLI).

### Configuring Automatic Failover

- **Note:** Before you begin configuring automatic failover, you must shut down your cluster. It is not currently possible to transition from a manual failover setup to an automatic failover setup while the cluster is running.

Configuring automatic failover requires two additional configuration parameters. In your `hdfs-site.xml` file, add:

```
<property>
  <name>dfs.ha.automatic-failover.enabled</name>
  <value>true</value>
</property>
```

This specifies that the cluster should be set up for automatic failover. In your `core-site.xml` file, add:

```
<property>
  <name>ha.zookeeper.quorum</name>
  <value>zk1.example.com:2181,zk2.example.com:2181,zk3.example.com:2181</value>
</property>
```

This lists the host-port pairs running the ZooKeeper service.

As with the parameters described earlier in this document, these settings may be configured on a per-nameservice basis by suffixing the configuration key with the nameservice ID. For example, in a cluster with federation enabled, you can explicitly enable automatic failover for only one of the nameservices by setting `dfs.ha.automatic-failover.enabled.my-nameservice-id`.

There are several other configuration parameters which you can set to control the behavior of automatic failover, but they are not necessary for most installations. See the configuration section of the [Hadoop documentation](#) for details.

### Initializing the HA state in ZooKeeper

After you have added the configuration keys, the next step is to initialize the required state in ZooKeeper. You can do so by running the following command from one of the NameNode hosts.

- **Note:** The ZooKeeper ensemble must be running when you use this command; otherwise it will not work properly.

```
$ hdfs zkfc -formatZK
```

This will create a `znode` in ZooKeeper in which the automatic failover system stores its data.

### Securing access to ZooKeeper

If you are running a secure cluster, you will probably want to ensure that the information stored in ZooKeeper is also secured. This prevents malicious clients from modifying the metadata in ZooKeeper or potentially triggering a false failover.

In order to secure the information in ZooKeeper, first add the following to your `core-site.xml` file:

```
<property>
  <name>ha.zookeeper.auth</name>
  <value>@/path/to/zk-auth.txt</value>
</property>
<property>
  <name>ha.zookeeper.acl</name>
  <value>@/path/to/zk-acl.txt</value>
</property>
```

Note the '@' character in these values – this specifies that the configurations are not inline, but rather point to a file on disk.

The first configured file specifies a list of ZooKeeper authentications, in the same format as used by the ZooKeeper CLI. For example, you may specify something like `digest:hdfs-zkfc:mypassword` where `hdfs-zkfc` is a unique username for ZooKeeper, and `mypassword` is some unique string used as a password.

Next, generate a ZooKeeper Access Control List (ACL) that corresponds to this authentication, using a command such as the following:

```
$ java -cp $ZK_HOME/lib/*:$ZK_HOME/zookeeper-3.4.2.jar
org.apache.zookeeper.server.auth.DigestAuthenticationProvider hdfs-zkfc:mypassword
output: hdfs-zkfc:mypassword->hdfs-zkfc:P/OQvnYyU/nF/mGYvB/xurX8dYs=
```

Copy and paste the section of this output after the '-'> string into the file `zk-acls.txt`, prefixed by the string "digest:". For example:

```
digest:hdfs-zkfc:vlUvLnd8MlacsE80rDuu6ONESbM=:rwcda
```

To put these ACLs into effect, rerun the `zkfc -formatZK` command as described above.

After doing so, you can verify the ACLs from the ZooKeeper CLI as follows:

```
[zk: localhost:2181(CONNECTED) 1] getAcl /hadoop-ha
'digest, 'hdfs-zkfc:vlUvLnd8MlacsE80rDuu6ONESbM=
: cdrwa
```

### Automatic Failover FAQ

#### Is it important that I start the ZKFC and NameNode daemons in any particular order?

No. On any given node you may start the ZKFC before or after its corresponding NameNode.

#### What additional monitoring should I put in place?

You should add monitoring on each host that runs a NameNode to ensure that the ZKFC remains running. In some types of ZooKeeper failures, for example, the ZKFC may unexpectedly exit, and should be restarted.

to ensure that the system is ready for automatic failover. Additionally, you should monitor each of the servers in the ZooKeeper quorum. If ZooKeeper crashes, automatic failover will not function.

### What happens if ZooKeeper goes down?

If the ZooKeeper cluster crashes, no automatic failovers will be triggered. However, HDFS will continue to run without any impact. When ZooKeeper is restarted, HDFS will reconnect with no issues.

### Can I designate one of my NameNodes as primary/preferred?

No. Currently, this is not supported. Whichever NameNode is started first will become active. You may choose to start the cluster in a specific order such that your preferred node starts first.

### How can I initiate a manual failover when automatic failover is configured?

Even if automatic failover is configured, you can initiate a [manual failover](#). It will perform a coordinated failover.

### Deploying HDFS High Availability

After you have set all of the necessary configuration options, you are ready to start the JournalNodes and the two HA NameNodes.

- **Important: Before you start:** Make sure you have performed all the configuration and setup tasks described under [Configuring Hardware for HDFS HA](#) on page 213 and [Configuring Software for HDFS HA](#) on page 216, including initializing the HA state in ZooKeeper if you are deploying automatic failover.

### Install and Start the JournalNodes

1. Install the JournalNode daemons on each of the machines where they will run.

#### To install JournalNode on Red Hat-compatible systems:

```
$ sudo yum install hadoop-hdfs-journalnode
```

#### To install JournalNode on Ubuntu and Debian systems:

```
$ sudo apt-get install hadoop-hdfs-journalnode
```

#### To install JournalNode on SLES systems:

```
$ sudo zypper install hadoop-hdfs-journalnode
```

2. Start the JournalNode daemons on each of the machines where they will run:

```
sudo service hadoop-hdfs-journalnode start
```

Wait for the daemons to start before formatting the primary NameNode (in a new cluster) and before starting the NameNodes (in all cases).

### Format the NameNode (if new cluster)

If you are setting up a new HDFS cluster, format the NameNode you will use as your primary NameNode; see [Formatting the NameNode](#).

- **Important:** Make sure the JournalNodes have started. Formatting will fail if you have configured the NameNode to communicate with the JournalNodes, but have not started the JournalNodes.



*Initialize the Shared Edits directory (if converting existing non-HA cluster)*

If you are converting a non-HA NameNode to HA, initialize the shared edits directory with the edits data from the local NameNode edits directories:

```
hdfs namenode -initializeSharedEdits
```

*Start the NameNodes*

1. Start the primary (formatted) NameNode:

```
$ sudo service hadoop-hdfs-namenode start
```

2. Start the standby NameNode:

```
$ sudo -u hdfs hdfs namenode -bootstrapStandby
$ sudo service hadoop-hdfs-namenode start
```

- **Note:** If [Kerberos is enabled](#), do not use commands in the form `sudo -u <user> <command>`; they will fail with a security error. Instead, use the following commands: `$ kinit <user>` (if you are using a password) *or* `$ kinit -kt <keytab> <principal>` (if you are using a keytab) and then, for each command executed by this user, `$ <command>`

Starting the standby NameNode with the `-bootstrapStandby` option copies over the contents of the primary NameNode's metadata directories (including the namespace information and most recent checkpoint) to the standby NameNode. (The location of the directories containing the NameNode metadata is configured via the configuration options `dfs.namenode.name.dir` and/or `dfs.namenode.edits.dir`.)

You can visit each NameNode's web page by browsing to its configured HTTP address. Notice that next to the configured address is the HA state of the NameNode (either "Standby" or "Active".) Whenever an HA NameNode starts and automatic failover is not enabled, it is initially in the Standby state. If automatic failover is enabled the first NameNode that is started will become active.

*Restart Services (if converting existing non-HA cluster)*

If you are converting from a non-HA to an HA configuration, you need to restart the JobTracker and TaskTracker (for MRv1, if used), or ResourceManager, NodeManager, and JobHistory Server (for YARN), and the DataNodes:

On each DataNode:

```
$ sudo service hadoop-hdfs-datanode start
```

On each TaskTracker system (MRv1):

```
$ sudo service hadoop-0.20-mapreduce-tasktracker start
```

On the JobTracker system (MRv1):

```
$ sudo service hadoop-0.20-mapreduce-jobtracker start
```

Verify that the JobTracker and TaskTracker started properly:

```
sudo jps | grep Tracker
```

On the ResourceManager system (YARN):

```
$ sudo service hadoop-yarn-resourcemanager start
```

On each NodeManager system (YARN; typically the same ones where DataNode service runs):

```
$ sudo service hadoop-yarn-nodemanager start
```

On the MapReduce JobHistory Server system (YARN):

```
$ sudo service hadoop-mapreduce-historyserver start
```

### *Deploy Automatic Failover (if it is configured)*

If you have configured automatic failover using the ZooKeeper FailoverController (ZKFC), you must install and start the `zkfc` daemon on each of the machines that runs a NameNode. Proceed as follows.

#### **To install ZKFC on Red Hat-compatible systems:**

```
$ sudo yum install hadoop-hdfs-zkfc
```

#### **To install ZKFC on Ubuntu and Debian systems:**

```
$ sudo apt-get install hadoop-hdfs-zkfc
```

#### **To install ZKFC on SLES systems:**

```
$ sudo zypper install hadoop-hdfs-zkfc
```

#### **To start the `zkfc` daemon:**

```
$ sudo service hadoop-hdfs-zkfc start
```

It is not important that you start the ZKFC and NameNode daemons in a particular order. On any given node you can start the ZKFC before or after its corresponding NameNode.

You should add monitoring on each host that runs a NameNode to ensure that the ZKFC remains running. In some types of ZooKeeper failures, for example, the ZKFC may unexpectedly exit, and should be restarted to ensure that the system is ready for automatic failover.

Additionally, you should monitor each of the servers in the ZooKeeper quorum. If ZooKeeper crashes, then automatic failover will not function. If the ZooKeeper cluster crashes, no automatic failovers will be triggered. However, HDFS will continue to run without any impact. When ZooKeeper is restarted, HDFS will reconnect with no issues.

### *Verifying Automatic Failover*

After the initial deployment of a cluster with automatic failover enabled, you should test its operation. To do so, first locate the active NameNode. As mentioned above, you can tell which node is active by visiting the NameNode web interfaces.

Once you have located your active NameNode, you can cause a failure on that node. For example, you can use `kill -9 <pid of NN>` to simulate a JVM crash. Or you can power-cycle the machine or its network interface to simulate different kinds of outages. After you trigger the outage you want to test, the other NameNode should automatically become active within several seconds. The amount of time required to detect a failure and trigger a failover depends on the configuration of `ha.zookeeper.session-timeout.ms`, but defaults to 5 seconds.

If the test does not succeed, you may have a misconfiguration. Check the logs for the `zkfc` daemons as well as the NameNode daemons in order to further diagnose the issue.

## Upgrading an HDFS HA Configuration to the Latest Release

### Upgrading to CDH 5

- **Important:** NFS shared storage is not supported in CDH 5. If you are using an HDFS HA configuration using NFS shared storage, [disable the configuration](#) before you begin the upgrade. You can [redploy HA](#) using Quorum-based storage either before or after the [upgrade](#).

To upgrade an HDFS HA configuration using Quorum-base storage from CDH 4 to the latest release, follow the directions for upgrading a cluster under [Upgrading from CDH 4 to CDH 5](#).

## Disabling and Redeploying HDFS HA

### Disabling and Redeploying HDFS HA Using Cloudera Manager

**Required Role:** [Cluster Administrator](#) [Full Administrator](#)

1. Go to the HDFS service.
2. Select **Actions** > **Disable High Availability**.
3. Select the hosts for the NameNode and the SecondaryNameNode and click **Continue**.
4. Select the HDFS checkpoint directory and click **Continue**.
5. Confirm that you want to take this action.
6. [Update the Hive Metastore NameNode](#).

Cloudera Manager ensures that one NameNode is active, and saves the namespace. Then it stops the standby NameNode, creates a SecondaryNameNode, removes the standby NameNode role, and restarts all the HDFS services.

### Disabling and Redeploying HDFS HA Using the Command Line

If you need to unconfigure HA and revert to using a single NameNode, either permanently or for [upgrade](#) or testing purposes, proceed as follows.

- **Important:** If you have been using NFS shared storage in CDH 4, you must unconfigure it before upgrading to CDH 5. Only [Quorum-based storage](#) is supported in CDH 5. If you already using Quorum-based storage, you *do not* need to unconfigure it in order to upgrade.

#### Step 1: Shut Down the Cluster

1. Shut down Hadoop services across your entire cluster. Do this from Cloudera Manager; or, if you are not using Cloudera Manager, run the following command on every host in your cluster:

```
$ for x in `cd /etc/init.d ; ls hadoop-*` ; do sudo service $x stop ; done
```

2. Check each host to make sure that there are no processes running as the `hdfs`, `yarn`, `mapred` or `httpfs` users from root:

```
# ps -aef | grep java
```

#### Step 2: Unconfigure HA

1. Disable the software configuration.
  - If you are using Quorum-based storage and want to unconfigure it, unconfigure the HA properties described under [Enabling HDFS HA Using the Command Line](#) on page 216.

If you intend to redeploy HDFS HA later, comment out the HA properties rather than deleting them.

- If you were using NFS shared storage in CDH 4, you must unconfigure the properties described below before upgrading to CDH 5.
2. Move the NameNode metadata directories on the standby NameNode. The location of these directories is configured by `dfs.namenode.name.dir` and/or `dfs.namenode.edits.dir`. Move them to a backup location.

### Step 3: Restart the Cluster

```
for x in `cd /etc/init.d ; ls hadoop-*` ; do sudo service $x start ; done
```

### Properties to unconfigure to disable an HDFS HA configuration using NFS shared storage

- **Important:** HDFS HA with NFS shared storage is not supported in CDH 5. Comment out or delete these properties before attempting to upgrade your cluster to CDH 5. (If you intend to configure HA with Quorum-based storage under CDH 5, you should comment them out rather than deleting them, as they are also used in that configuration.)

Unconfigure the following properties:

- In your `core-site.xml` file:

**fs.defaultFS** (formerly `fs.default.name`)

Optionally, you may have configured the default path for Hadoop clients to use the HA-enabled logical URI. For example, if you used `mycluster` as the nameservice ID as shown below, this will be the value of the authority portion of all of your HDFS paths.

```
<property>
  <name>fs.default.name</name>
  <value>hdfs://mycluster</value>
</property>
```

- In your `hdfs-site.xml` configuration file:

**dfs.nameservices**

```
<property>
  <name>dfs.nameservices</name>
  <value>mycluster</value>
</property>
```

- **Note:** If you are also using HDFS federation, this configuration setting will include the list of other nameservices, HA or otherwise, as a comma-separated list.

**dfs.ha.namenodes.[nameservice ID]**

A list of comma-separated NameNode IDs used by DataNodes to determine all the NameNodes in the cluster. For example, if you used `mycluster` as the nameservice ID, and you used `nn1` and `nn2` as the individual IDs of the NameNodes, you would have configured this as follows:

```
<property>
  <name>dfs.ha.namenodes.mycluster</name>
  <value>nn1,nn2</value>
</property>
```

**dfs.namenode.rpc-address.[nameservice ID]**

For both of the previously-configured NameNode IDs, the full address and RPC port of the NameNode process. For example:

```
<property>
  <name>dfs.namenode.rpc-address.mycluster.nn1</name>
  <value>machine1.example.com:8020</value>
</property>
<property>
  <name>dfs.namenode.rpc-address.mycluster.nn2</name>
  <value>machine2.example.com:8020</value>
</property>
```

- **Note:** You may have similarly configured the `servicerpc-address` setting.

#### **dfs.namenode.http-address.[nameservice ID]**

The addresses for both NameNodes' HTTP servers to listen on. For example:

```
<property>
  <name>dfs.namenode.http-address.mycluster.nn1</name>
  <value>machine1.example.com:50070</value>
</property>
<property>
  <name>dfs.namenode.http-address.mycluster.nn2</name>
  <value>machine2.example.com:50070</value>
</property>
```

- **Note:** If you have Hadoop's Kerberos security features enabled, and you use HSFTP, you will have set the `https-address` similarly for each NameNode.

#### **dfs.namenode.shared.edits.dir**

The path to the remote shared edits directory which the standby NameNode uses to stay up-to-date with all the file system changes the Active NameNode makes. You should have configured only one of these directories, mounted read/write on both NameNode machines. The value of this setting should be the absolute path to this directory on the NameNode machines. For example:

```
<property>
  <name>dfs.namenode.shared.edits.dir</name>
  <value>file:///mnt/filer1/dfs/ha-name-dir-shared</value>
</property>
```

#### **dfs.client.failover.proxy.provider.[nameservice ID]**

The name of the Java class that the DFS client uses to determine which NameNode is the current active, and therefore which NameNode is currently serving client requests. The only implementation which shipped with Hadoop is the `ConfiguredFailoverProxyProvider`. For example:

```
<property>
  <name>dfs.client.failover.proxy.provider.mycluster</name>

  <value>org.apache.hadoop.hdfs.server.namenode.ha.ConfiguredFailoverProxyProvider</value>
</property>
```

**dfs.ha.fencing.methods** - a list of scripts or Java classes which will be used to fence the active NameNode during a failover.

- **Note:** If you implemented your own custom fencing method, see the `org.apache.hadoop.ha.NodeFencer` class.

- **The sshfence fencing method**

`sshfence` - SSH to the active NameNode and kill the process

For example:

```
<property>
  <name>dfs.ha.fencing.methods</name>
  <value>sshfence</value>
</property>

<property>
  <name>dfs.ha.fencing.ssh.private-key-files</name>
  <value>/home/exampleuser/.ssh/id_rsa</value>
</property>
```

Optionally, you may have configured a non-standard username or port to perform the SSH, as shown below, and also a timeout, in milliseconds, for the SSH:

```
<property>
  <name>dfs.ha.fencing.methods</name>
  <value>sshfence([[username]][ :port]])</value>
</property>
<property>
  <name>dfs.ha.fencing.ssh.connect-timeout</name>
  <value>30000</value>
  <description>
    SSH connection timeout, in milliseconds, to use with the builtin
    sshfence fencer.
  </description>
</property>
```

- **The shell fencing method**

`shell` - run an arbitrary shell command to fence the active NameNode

The shell fencing method runs an arbitrary shell command, which you may have configured as shown below:

```
<property>
  <name>dfs.ha.fencing.methods</name>
  <value>shell(/path/to/my/script.sh arg1 arg2 ...)</value>
</property>
```

**Automatic failover:** If you configured automatic failover, you configured two additional configuration parameters.

- In your `hdfs-site.xml`:

```
<property>
  <name>dfs.ha.automatic-failover.enabled</name>
  <value>true</value>
</property>
```

- In your `core-site.xml` file, add:

```
<property>
  <name>ha.zookeeper.quorum</name>
  <value>zk1.example.com:2181,zk2.example.com:2181,zk3.example.com:2181</value>
</property>
```

**Other properties:** There are several other configuration parameters which you may have set to control the behavior of automatic failover, though they were not necessary for most installations. See the configuration section of the [Hadoop documentation](#) for details.

## Redeploying HDFS High Availability

If you need to redeploy HA using Quorum-based storage after temporarily disabling it, proceed as follows:

1. Shut down the cluster as described in [Step 1: Shut Down the Cluster](#) on page 227.

2. Uncomment the properties you commented out in [Step 2: Unconfigure HA](#) on page 227.
3. Deploy HDFS HA, following the instructions under [Deploying HDFS High Availability](#) on page 224.

## Configuring Other CDH Components to Use HDFS HA

You can use the HDFS high availability NameNodes with other components of CDH.

### Configuring HBase to Use HDFS HA

#### Configuring HBase to Use HDFS HA Using the Command Line

To configure HBase to use HDFS HA, proceed as follows.

#### Shut Down the HBase Cluster

1. Stop the Thrift server and clients:

```
sudo service hbase-thrift stop
```

2. Stop the cluster by shutting down the Master and the RegionServers:

- Use the following command on the Master host:

```
sudo service hbase-master stop
```

- Use the following command on each host hosting a RegionServer:

```
sudo service hbase-regionserver stop
```

#### Configure hbase.rootdir

Change the distributed file system URI in `hbase-site.xml` to the name specified in the `dfs.nameservices` property in `hdfs-site.xml`. The clients must also have access to `hdfs-site.xml`'s `dfs.client.*` settings to properly use HA.

For example, suppose the HDFS HA property `dfs.nameservices` is set to `ha-nn` in `hdfs-site.xml`. To configure HBase to use the HA NameNodes, specify that same value as part of your `hbase-site.xml`'s `hbase.rootdir` value:

```
<!-- Configure HBase to use the HA NameNode nameservice -->
<property>
  <name>hbase.rootdir</name>
  <value>hdfs://ha-nn/hbase</value>
</property>
```

#### Restart HBase

1. Start the HBase Master.
2. Start each of the HBase RegionServers.

#### HBase-HDFS HA Troubleshooting

Problem: HMasters fail to start.

Solution: Check for this error in the HMaster log:

```
2012-05-17 12:21:28,929 FATAL master.HMaster (HMaster.java:abort(1317)) - Unhandled
exception. Starting shutdown.
java.lang.IllegalArgumentException: java.net.UnknownHostException: ha-nn
    at
```

```
org.apache.hadoop.security.SecurityUtil.buildTokenService(SecurityUtil.java:431)
    at
org.apache.hadoop.hdfs.NameNodeProxies.createNonHAProxy(NameNodeProxies.java:161)
    at org.apache.hadoop.hdfs.NameNodeProxies.createProxy(NameNodeProxies.java:126)
    ...
```

If so, verify that Hadoop's `hdfs-site.xml` and `core-site.xml` files are in your `hbase/conf` directory. This may be necessary if you put your configurations in non-standard places.

### Upgrading the Hive Metastore to Use HDFS HA

The Hive metastore can be configured to use HDFS high availability.

#### Upgrading the Hive Metastore to Use HDFS HA Using Cloudera Manager

1. Go the Hive service.
2. Select **Actions** > **Stop**.

- **Note:** You may want to stop the Hue and Impala services first, if present, as they depend on the Hive service.

Click **Stop** to confirm the command.

3. Back up the Hive metastore database.
4. Select **Actions** > **Update Hive Metastore NameNodes** and confirm the command.
5. Select **Actions** > **Start**.
6. Restart the Hue and Impala services if you stopped them prior to updating the metastore.

#### Upgrading the Hive Metastore to Use HDFS HA Using the Command Line

To configure the Hive metastore to use HDFS HA, change the records to reflect the location specified in the `dfs.nameservices` property, using the `Hive metatool` to obtain and change the locations.

- **Note:** Before attempting to upgrade the Hive metastore to use HDFS HA, shut down the metastore and back it up to a persistent store.

If you are unsure which version of Avro SerDe is used, use both the `serdePropKey` and `tablePropKey` arguments. For example:

```
$ metatool -listFSRoot
hdfs://oldnamenode.com/user/hive/warehouse
$ metatool -updateLocation hdfs://nameservice1 hdfs://oldnamenode.com -tablePropKey
avro.schema.url
-serdePropKey schema.url
$ metatool -listFSRoot
hdfs://nameservice1/user/hive/warehouse
```

where:

- `hdfs://oldnamenode.com/user/hive/warehouse` identifies the NameNode location.
- `hdfs://nameservice1` specifies the new location and should match the value of the `dfs.nameservices` property.
- `tablePropKey` is a table property key whose value field may reference the HDFS NameNode location and hence may require an update. To update the Avro SerDe schema URL, specify `avro.schema.url` for this argument.
- `serdePropKey` is a SerDe property key whose value field may reference the HDFS NameNode location and hence may require an update. To update the Haivvero schema URL, specify `schema.url` for this argument.

- **Note:** The `Hive metatool` is a best effort service that tries to update as many Hive metastore records as possible. If it encounters an error during the update of a record, it skips to the next record.



## Configuring Hue to Work with HDFS HA

1. Add the [HttpFS](#) role.
2. After the command has completed, go to the **Hue** service.
3. Click the **Configuration** tab.
4. Select the **Service-Wide > HDFS Web Interface Role** property.
5. Select **HttpFS** instead of the NameNode role, and save your changes.
6. Restart the Hue service.

## Configuring Impala to Work with HDFS HA

1. Complete the steps to reconfigure the Hive metastore database, as described in the preceding section. Impala shares the same underlying database with Hive, to manage metadata for databases, tables, and so on.
2. Issue the `INVALIDATE METADATA` statement from an Impala shell. This one-time operation makes all Impala daemons across the cluster aware of the latest settings for the Hive metastore database. Alternatively, restart the Impala service.

## Configuring Oozie to Use HDFS HA

To configure an Oozie workflow to use HDFS HA, use the HDFS nameservice instead of the NameNode URI in the `<name-node>` element of the workflow.

**Example:**

```
<action name="mr-node">
  <map-reduce>
    <job-tracker>${jobTracker}</job-tracker>
    <name-node>hdfs://ha-nn
```

where *ha-nn* is the value of `dfs.nameservices` in `hdfs-site.xml`.

## Administering an HDFS High Availability Cluster

### Manually Failing Over to the Standby NameNode

#### Using Cloudera Manager

If you are running a HDFS service with HA enabled, you can manually cause the active NameNode to failover to the standby NameNode. This is useful for planned downtime—for hardware changes, configuration changes, or software upgrades of your primary host.

1. Go to the HDFS service.
2. Click the **Instances** tab.
3. Select **Actions > Manual Failover**. (This option does not appear if HA is not enabled for the cluster.)
4. From the pop-up, select the NameNode that should be made active, then click **Manual Failover**.

- **Note: For advanced use only:** You can set the **Force Failover** checkbox to force the selected NameNode to be active, irrespective of its state or the other NameNode's state. Forcing a failover will first attempt to failover the selected NameNode to active mode and the other NameNode to standby mode. It will do so even if the selected NameNode is in safe mode. If this fails, it will proceed to transition the selected NameNode to active mode. To avoid having two NameNodes be active, use this only if the other NameNode is either definitely stopped, or can be transitioned to standby mode by the first failover step.

5. When all the steps have been completed, click **Finish**.

Cloudera Manager transitions the NameNode you selected to be the active NameNode, and the other NameNode to be the standby NameNode. HDFS should *never* have two active NameNodes.

### Using the Command Line

To initiate a failover between two NameNodes, run the command `hdfs haadmin - failover`.

This command causes a failover from the first provided NameNode to the second. If the first NameNode is in the Standby state, this command simply transitions the second to the Active state without error. If the first NameNode is in the Active state, an attempt will be made to gracefully transition it to the Standby state. If this fails, the fencing methods (as configured by `dfs.ha.fencing.methods`) will be attempted in order until one of the methods succeeds. Only after this process will the second NameNode be transitioned to the Active state. If no fencing method succeeds, the second NameNode will not be transitioned to the Active state, and an error will be returned.

- **Note:** Running `hdfs haadmin -failover` from the command line works whether you have configured HA from the command line or via Cloudera Manager. This means you can initiate a failover manually even if Cloudera Manager is unavailable.

### Other hdfs haadmin Commands

After your HA NameNodes are configured and started, you will have access to some additional commands to administer your HA HDFS cluster. Specifically, you should familiarize yourself with the subcommands of the `hdfs haadmin` command.

This page describes high-level uses of some important subcommands. For specific usage information of each subcommand, you should run `hdfs haadmin -help <command>`.

#### getServiceState

`getServiceState` - determine whether the given NameNode is Active or Standby

Connect to the provided NameNode to determine its current state, printing either "standby" or "active" to `STDOUT` as appropriate. This subcommand might be used by `cron` jobs or monitoring scripts which need to behave differently based on whether the NameNode is currently Active or Standby.

#### checkHealth

`checkHealth` - check the health of the given NameNode

Connect to the provided NameNode to check its health. The NameNode is capable of performing some diagnostics on itself, including checking if internal services are running as expected. This command will return 0 if the NameNode is healthy, non-zero otherwise. One might use this command for monitoring purposes.

- **Note:** The `checkHealth` command is not yet implemented, and at present will always return success, unless the given NameNode is completely down.

### Using the dfsadmin command when HA is enabled

In previous versions of Hadoop, when HA was enabled, the `dfsadmin` command would not run operations on both active and standby NameNodes by default, even if the operations were permitted to run on both active and standby NameNodes. Due to an enhancement introduced in [HDFS-6507](#) (included in CDH 5.2), appropriate operations, such as `-refreshNodes`, `-refreshServiceAcl`, `-refreshUserToGroupsMappings`, and `-refreshSuperUserGroupsConfiguration`, now run on both active and standby NameNodes, unless you use the `-fs` option to specify a specific NameNode on which to run the operations.

### Moving an HA NameNode to a New Host

#### Using the Command Line

Use the following steps to move one of the NameNodes to a new host.

In this example, the current NameNodes are called `nn1` and `nn2`, and the new NameNode is `nn2-alt`. The example assumes that `nn2-alt` is already a member of this CDH 5 HA cluster, that automatic failover is [configured](#) and that a JournalNode on `nn2` is to be moved to `nn2-alt`, in addition to NameNode service itself.

The procedure moves the NameNode and JournalNode services from `nn2` to `nn2-alt`, reconfigures `nn1` to recognize the new location of the JournalNode, and restarts `nn1` and `nn2-alt` in the new HA configuration.

### Step 1: Make sure that `nn1` is the active NameNode

Make sure that the NameNode that is *not* going to be moved is active; in this example, `nn1` must be active. You can use the NameNodes' web UIs to see which is active; see [Start the NameNodes](#) on page 225.

If `nn1` is not the active NameNode, use the `hdfs haadmin -failover` command to initiate a failover from `nn2` to `nn1`:

```
hdfs haadmin -failover nn2 nn1
```

### Step 2: Stop services on `nn2`

Once you've made sure that the node to be moved is inactive, stop services on that node: in this example, stop services on `nn2`. Stop the NameNode, the ZKFC daemon if this an automatic-failover deployment, and the JournalNode if you are moving it. Proceed as follows.

1. Stop the NameNode daemon:

```
$ sudo service hadoop-hdfs-namenode stop
```

2. Stop the ZKFC daemon if it is running:

```
$ sudo service hadoop-hdfs-zkfc stop
```

3. Stop the JournalNode daemon if it is running:

```
$ sudo service hadoop-hdfs-journalnode stop
```

4. Make sure these services are not set to restart on boot. If you are not planning to use `nn2` as a NameNode again, you may want remove the services.

### Step 3: Install the NameNode daemon on `nn2-alt`

See the instructions for installing `hadoop-hdfs-namenode` in the *CDH 5 Installation Guide* under [Step 3: Install CDH 5 with YARN](#) or [Step 4: Install CDH 5 with MRv1](#).

### Step 4: Configure HA on `nn2-alt`

See [Enabling HDFS HA](#) on page 213 for the properties to configure on `nn2-alt` in `core-site.xml` and `hdfs-site.xml`, and explanations and instructions. You should copy the values that are already set in the corresponding files on `nn2`.

- If you are relocating a JournalNode to `nn2-alt`, follow [these directions](#) to install it, but don't start it yet.
- If you are using automatic failover, make sure you follow the [instructions](#) for configuring the necessary properties on `nn2-alt` and initializing the HA state in Zookeeper.

- **Note:** You do not need to shut down the cluster to do this if automatic failover is already configured as your failover method; shutdown is required only if you are switching from manual to automatic failover.

### Step 5: Copy the contents of the `dfs.name.dir` and `dfs.journalnode.edits.dir` directories to `nn2-alt`

Use `rsync` or a similar tool to copy the contents of the `dfs.name.dir` directory, and the `dfs.journalnode.edits.dir` directory if you are moving the `JournalNode`, from `nn2` to `nn2-alt`.

### Step 6: If you are moving a `JournalNode`, update `dfs.namenode.shared.edits.dir` on `nn1`

If you are relocating a `JournalNode` from `nn2` to `nn2-alt`, update `dfs.namenode.shared.edits.dir` in `hdfs-site.xml` on `nn1` to reflect the new hostname. See [this section](#) for more information about `dfs.namenode.shared.edits.dir`.

### Step 7: If you are using automatic failover, install the `zkfc` daemon on `nn2-alt`

For instructions, see [Deploy Automatic Failover \(if it is configured\)](#) on page 226, but do not start the daemon yet.

### Step 8: Start services on `nn2-alt`

Start the `NameNode`; start the `ZKFC` for automatic failover; and install and start a `JournalNode` if you want one to run on `nn2-alt`. Proceed as follows.

1. Start the `JournalNode` daemon:

```
$ sudo service hadoop-hdfs-journalnode start
```

2. Start the `NameNode` daemon:

```
$ sudo service hadoop-hdfs-namenode start
```

3. Start the `ZKFC` daemon:

```
$ sudo service hadoop-hdfs-zkfc start
```

4. Set these services to restart on boot; for example on a RHEL-compatible system:

```
$ sudo chkconfig hadoop-hdfs-namenode on
$ sudo chkconfig hadoop-hdfs-zkfc on
$ sudo chkconfig hadoop-hdfs-journalnode on
```

### Step 9: If you are relocating a `JournalNode`, fail over to `nn2-alt`

```
hdfs haadmin -failover nn1 nn2-alt
```

### Step 10: If you are relocating a `JournalNode`, restart `nn1`

Restart the `NameNode` daemon on `nn1` to force it to re-read the configuration:

```
$ sudo service hadoop-hdfs-namenode stop
$ sudo service hadoop-hdfs-namenode start
```

## Converting From an NFS-mounted Shared Edits Directory to Quorum-based Storage Using Cloudera Manager

Converting a HA configuration from using an NFS-mounted shared edits directory to Quorum-based storage involves disabling the current HA configuration then enabling HA using Quorum-based storage.

1. [Disable HA](#).
2. Although the standby `NameNode` role is removed, its name directories are not deleted. Empty these directories.
3. [Enable HA with Quorum-based storage](#).

### Using the Command Line

To switch from shared storage using NFS to Quorum-based storage, proceed as follows:

1. [Disable HA.](#)
2. [Redeploy HA using Quorum-based storage.](#)

## MapReduce (MRv1) and YARN (MRv2) High Availability

This section covers:

### YARN (MRv2) ResourceManager High Availability

The YARN ResourceManager (RM) is responsible for tracking the resources in a cluster and scheduling applications (for example, MapReduce jobs). Before CDH 5, the RM was a single point of failure in a YARN cluster. The RM high availability (HA) feature adds redundancy in the form of an Active/Standby RM pair to remove this single point of failure. Furthermore, upon failover from the Standby RM to the Active, the applications can resume from their last check-pointed state; for example, completed map tasks in a MapReduce job are not re-run on a subsequent attempt. This allows events such the following to be handled without any significant performance effect on running applications.:

- Unplanned events such as machine crashes
- Planned maintenance events such as software or hardware upgrades on the machine running the ResourceManager.

RM HA requires ZooKeeper and HDFS services to be running.

### Architecture

RM HA is implemented by means of an active-standby pair of RMs. On start-up, each RM is in the standby state: the process is started, but the state is not loaded. When transitioning to active, the RM loads the internal state from the designated state store and starts all the internal services. The stimulus to transition-to-active comes from either the administrator (through the [CLI](#)) or through the integrated failover controller when [automatic failover](#) is enabled. The subsections that follow provide more details about the components of RM HA.

#### RM Restart

RM restart allows restarting the RM, while recovering the in-flight applications if recovery is enabled. To achieve this, the RM stores its internal state, primarily application-related data and tokens, to the `RMStateStore`; the cluster resources are re-constructed when the NodeManagers connect. The available alternatives for the state store are `MemoryRMStateStore` (a memory-based implementation), `FileSystemRMStateStore` (file system-based implementation; HDFS can be used for the file system), and `ZKRMStateStore` (ZooKeeper-based implementation).

#### Fencing

When running two RMs, a split-brain situation can arise where both RMs assume they are Active. To avoid this, only a single RM should be able to perform active operations and the other RM should be "fenced". The ZooKeeper-based state store (`ZKRMStateStore`) allows a single RM to make changes to the stored state, implicitly fencing the other RM. This is accomplished by the RM claiming exclusive create-delete permissions on the root `znode`. The ACLs on the root `znode` are automatically created based on the ACLs configured for the store; in case of secure clusters, Cloudera recommends that you set ACLs for the root node such that both RMs share read-write-admin access, but have exclusive create-delete access. The fencing is implicit and doesn't require explicit configuration (as fencing in HDFS and MRv1 does). You can plug in a custom "Fencer" if you choose to – for example, to use a different implementation of the state store.

#### Configuration and FailoverProxy

In an HA setting, you should configure two RMs to use different ports (for example, ports on different hosts). To facilitate this, YARN uses the notion of an RM Identifier (`rm-id`). Each RM has a unique `rm-id`, and all the RPC configurations (`<rpc-address>`; for example `yarn.resourcemanager.address`) for that RM can be configured

## High Availability

via `<rpc-address>.<rm-id>`. Clients, ApplicationMasters, and NodeManagers use these RPC addresses to talk to the active RM automatically, even after a failover. To achieve this, they cycle through the list of RMs in the configuration. This is done automatically and doesn't require any configuration (as it does in HDFS and MapReduce (MRv1)).

### Automatic Failover

By default, RM HA uses ZKFC (ZooKeeper-based failover controller) for automatic failover in case the active RM is unreachable or goes down. Internally, the **ActiveStandbyElector** is used to elect the Active RM. The failover controller runs as part of the RM (not as a separate process as in HDFS and MapReduce v1) and requires no further setup after the appropriate properties are [configured](#) in `yarn-site.xml`.

You can plug in a custom failover controller if you prefer.

### Manual Transitions and Failover

You can use the [command-line tool](#) `yarn rmadmin` to transition a particular RM to active or standby state, to fail over from one RM to the other, to get the HA state of an RM, and to monitor an RM's health.

## Configuring YARN (MRv2) ResourceManager High Availability Using Cloudera Manager

**Required Role:** Cluster Administrator Full Administrator

You can use Cloudera Manager to configure CDH 5 or later for ResourceManager high availability (HA). Cloudera Manager supports automatic failover of the ResourceManager. It does not provide a mechanism to manually force a failover through the Cloudera Manager user interface.

- **Important:** Enabling or disabling HA will cause the previous monitoring history to become unavailable.

### Enabling High Availability

1. Go to the YARN service.
2. Select **Actions > Enable High Availability**. A screen showing the hosts that are eligible to run a standby ResourceManager displays. The host where the current ResourceManager is running is not available as a choice.
3. Select the host where you want the standby ResourceManager to be installed, and click **Continue**. Cloudera Manager proceeds to execute a set of commands that stop the YARN service, add a standby ResourceManager, initialize the ResourceManager high availability state in ZooKeeper, restart YARN, and redeploy the relevant client configurations.
4. Work preserving recovery is enabled for the RM by default when you enable RM HA in Cloudera Manager. For more information, including instructions on disabling work preserving recovery, see [Work Preserving Recovery for YARN Components](#) on page 244.

- **Note:** ResourceManager HA doesn't affect the JobHistory Server (JHS). JHS doesn't maintain any state, so if the host fails you can simply assign it to a new host. You can also enable process auto-restart by doing the following:
  1. Go to the YARN service.
  2. Click the **Configuration** tab.
  3. Expand the **JobHistory Server Default Group**.
  4. Select the **Advanced** subcategory.
  5. Check the **Automatically Restart Process** checkbox.
  6. Restart the JobHistory Server role.

### Disabling High Availability

1. Go to the YARN service.
2. Select **Actions > Disable High Availability**. A screen showing the hosts running the ResourceManagers displays.

3. Select which ResourceManager (host) you want to remain as the single ResourceManager, and click **Continue**. Cloudera Manager executes a set of commands that stop the YARN service, remove the standby ResourceManager and the Failover Controller, restart the YARN service, and redeploy client configurations.

## Configuring YARN (MRv2) ResourceManager High Availability Using the Command Line

To configure and start ResourceManager HA, proceed as follows.

### Stop the YARN daemons

Stop the MapReduce JobHistory service, ResourceManager service, and NodeManager on all nodes where they are running, as follows:

```
$ sudo service hadoop-mapreduce-historyserver stop
$ sudo service hadoop-yarn-resourcemanager stop
$ sudo service hadoop-yarn-nodemanager stop
```

### Configure Manual Failover, and Optionally Automatic Failover

To configure failover:

- **Note:**

Configure the following properties in `yarn-site.xml` as shown, whether you are configuring manual or automatic failover. They are sufficient to configure manual failover. You need to configure additional properties for automatic failover.

Name	Used On	Default Value	Recommended Value	Description
<code>yarn.resourcemanager.ha.enabled</code>	ResourceManager, NodeManager, Client	false	true	Enable HA
<code>yarn.resourcemanager.ha.rm-ids</code>	ResourceManager, NodeManager, Client	(None)	Cluster-specific, e.g., <code>rm1,rm2</code>	Comma-separated list of ResourceManager ids in this cluster.
<code>yarn.resourcemanager.ha.id</code>	ResourceManager	(None)	RM-specific, e.g., <code>rm1</code>	Id of the current ResourceManager. Must be set explicitly on each ResourceManager to the appropriate value.
<code>yarn.resourcemanager.address.&lt;rm-id&gt;</code>	ResourceManager, Client	(None)	Cluster-specific	The value of <code>yarn.resourcemanager.address</code> (Client-RM RPC) for this RM. Must be set for all RMs.
<code>yarn.resourcemanager.scheduler.address.&lt;rm-id&gt;</code>	ResourceManager, Client	(None)	Cluster-specific	The value of <code>yarn.resourcemanager.scheduler.address</code> (AM-RM RPC) for this RM. Must be set for all RMs.

Name	Used On	Default Value	Recommended Value	Description
yarn.resourcemanager.admin.address.<rm-id>	ResourceManager, Client/Admin	(None)	Cluster-specific	The value of yarn.resourcemanager.admin.address (RM administration) for this RM. Must be set for all RMs.
yarn.resourcemanager.resource-tracker.address.<rm-id>	ResourceManager, NodeManager	(None)	Cluster-specific	The value of yarn.resourcemanager.resource-tracker.address (NM-RM RPC) for this RM. Must be set for all RMs.
yarn.resourcemanager.webapp.address.<rm-id>	ResourceManager, Client	(None)	Cluster-specific	The value of yarn.resourcemanager.webapp.address (RM webapp) for this RM. Must be set for all RMs.
yarn.resourcemanager.recovery.enabled	ResourceManager	false	true	Enable job recovery on RM restart or failover.
yarn.resourcemanager.store.class	ResourceManager	org.apache.hadoop.yarn.server.resourcemanager.recovery.FileSystemRMStateStore	org.apache.hadoop.yarn.server.resourcemanager.recovery.ZKRMStateStore	The RMStateStore implementation to use to store the ResourceManager's internal state. The ZooKeeper-based store supports fencing implicitly, i.e., allows a single ResourceManager to make multiple changes at a time, and hence is recommended.
yarn.resourcemanager.zk-address	ResourceManager	(None)	Cluster-specific	The ZooKeeper quorum to use to store the ResourceManager's internal state.
yarn.resourcemanager.zk-acl	ResourceManager	world:anyone:rwcd	Cluster-specific	The ACLs the ResourceManager uses for the znode structure to store the internal state.
yarn.resourcemanager.zk-state-store.root-node.acl	ResourceManager	(None)	Cluster-specific	The ACLs used for the root node of the ZooKeeper state store. The ACLs set here should allow



Name	Used On	Default Value	Recommended Value	Description
				both ResourceManagers to read, write, and administer, with exclusive access to create and delete. If nothing is specified, the root node ACLs are automatically generated on the basis of the ACLs specified through <code>yarn.resourcemanager.zk-acl</code> . But that leaves a security hole in a secure setup.

#### To configure automatic failover:

Configure the following additional properties in `yarn-site.xml` to configure automatic failover.

#### Configure work preserving recovery:

Optionally, you can configure work preserving recovery for the Resource Manager and Node Managers. See [Work Preserving Recovery for YARN Components](#) on page 244.

Name	Used On	Default Value	Recommended Value	Description
<code>yarn.resourcemanager.ha.automatic-failover.enabled</code>	ResourceManager	true	true	Enable automatic failover
<code>yarn.resourcemanager.ha.automatic-failover.enabled</code>	ResourceManager	true	true	Use the <code>EmbeddedElectorService</code> to pick an Active RM from the ensemble
<code>yarn.resourcemanager.cluster-id</code>	ResourceManager	No default value.	Cluster-specific	Cluster name used by the <code>ActiveStandbyElector</code> to elect one of the ResourceManagers as leader.

The following is a sample `yarn-site.xml` showing these properties configured, including work preserving recovery for both RM and NM:

```
<configuration>
<!-- Resource Manager Configs -->
  <property>
    <name>yarn.resourcemanager.connect.retry-interval.ms</name>
    <value>2000</value>
  </property>
  <property>
    <name>yarn.resourcemanager.ha.enabled</name>
    <value>true</value>
  </property>
  <property>
    <name>yarn.resourcemanager.ha.automatic-failover.enabled</name>
    <value>true</value>
```

```

</property>
<property>
  <name>yarn.resourcemanager.ha.automatic-failover.embedded</name>
  <value>true</value>
</property>
<property>
  <name>yarn.resourcemanager.cluster-id</name>
  <value>pseudo-yarn-rm-cluster</value>
</property>
<property>
  <name>yarn.resourcemanager.ha.rm-ids</name>
  <value>rm1,rm2</value>
</property>
<property>
  <name>yarn.resourcemanager.ha.id</name>
  <value>rm1</value>
</property>
<property>
  <name>yarn.resourcemanager.scheduler.class</name>
<value>org.apache.hadoop.yarn.server.resourcemanager.scheduler.fair.FairScheduler</value>
</property>
<property>
  <name>yarn.resourcemanager.recovery.enabled</name>
  <value>true</value>
</property>
<property>
  <name>yarn.resourcemanager.store.class</name>
  <value>org.apache.hadoop.yarn.server.resourcemanager.recovery.ZKRMStateStore</value>
</property>
<property>
  <name>yarn.resourcemanager.zk.state-store.address</name>
  <value>localhost:2181</value>
</property>
<property>
  <name>yarn.app.mapreduce.am.scheduler.connection.wait.interval-ms</name>
  <value>5000</value>
</property>
<property>
  <name>yarn.resourcemanager.work-preserving-recovery.enabled</name>
  <value>true</value>
</property>

<!-- RM1 configs -->
<property>
  <name>yarn.resourcemanager.address.rm1</name>
  <value>host1:23140</value>
</property>
<property>
  <name>yarn.resourcemanager.scheduler.address.rm1</name>
  <value>host1:23130</value>
</property>
<property>
  <name>yarn.resourcemanager.webapp.https.address.rm1</name>
  <value>host1:23189</value>
</property>
<property>
  <name>yarn.resourcemanager.webapp.address.rm1</name>
  <value>host1:23188</value>
</property>
<property>
  <name>yarn.resourcemanager.resource-tracker.address.rm1</name>
  <value>host1:23125</value>
</property>
<property>
  <name>yarn.resourcemanager.admin.address.rm1</name>
  <value>host1:23141</value>
</property>

<!-- RM2 configs -->
<property>

```

```

    <name>yarn.resourcemanager.address.rm2</name>
    <value>host2:23140</value>
  </property>
  <property>
    <name>yarn.resourcemanager.scheduler.address.rm2</name>
    <value>host2:23130</value>
  </property>
  <property>
    <name>yarn.resourcemanager.webapp.https.address.rm2</name>
    <value>host2:23189</value>
  </property>
  <property>
    <name>yarn.resourcemanager.webapp.address.rm2</name>
    <value>host2:23188</value>
  </property>
  <property>
    <name>yarn.resourcemanager.resource-tracker.address.rm2</name>
    <value>host2:23125</value>
  </property>
  <property>
    <name>yarn.resourcemanager.admin.address.rm2</name>
    <value>host2:23141</value>
  </property>
<!-- Node Manager Configs -->
  <property>
    <description>Address where the localizer IPC is.</description>
    <name>yarn.nodemanager.localizer.address</name>
    <value>0.0.0.0:23344</value>
  </property>
  <property>
    <description>NM Webapp address.</description>
    <name>yarn.nodemanager.webapp.address</name>
    <value>0.0.0.0:23999</value>
  </property>
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
  <property>
    <name>yarn.nodemanager.local-dirs</name>
    <value>/tmp/pseudo-dist/yarn/local</value>
  </property>
  <property>
    <name>yarn.nodemanager.log-dirs</name>
    <value>/tmp/pseudo-dist/yarn/log</value>
  </property>
  <property>
    <name>mapreduce.shuffle.port</name>
    <value>23080</value>
  </property>
  <property>
    <name>yarn.resourcemanager.work-preserving-recovery.enabled</name>
    <value>true</value>
  </property>
</configuration>

```

### Re-start the YARN daemons

Start the MapReduce JobHistory server, ResourceManager, and NodeManager on all nodes where they were previously running, as follows:

```

$ sudo service hadoop-mapreduce-historyserver start
$ sudo service hadoop-yarn-resourcemanager start
$ sudo service hadoop-yarn-nodemanager start

```

### Using `yarn rmadmin` to Administer ResourceManager HA

You can use `yarn rmadmin` on the command line to manage your ResourceManager HA deployment. `yarn rmadmin` has the following options related to RM HA:

```
[ -transitionToActive <serviceId> ]  
[ -transitionToStandby <serviceId> ]  
[ -getServiceState <serviceId> ]  
[ -checkHealth <serviceId> ]  
[ -help <command> ]
```

where *serviceId* is the `rm-id`.

- **Note:** Even though `-help` lists the `-failover` option, it is not supported by `yarn rmadmin`.

### Work Preserving Recovery for YARN Components

CDH 5.2 introduces *work preserving recovery* for the YARN ResourceManager and NodeManager. With work preserving recovery enabled, if a ResourceManager or NodeManager restarts, no in-flight work is lost. You can configure work preserving recovery separately for a ResourceManager or NodeManager.

- **Note:** YARN does not support high availability for the Job History Server (JHS). If the JHS goes down, Cloudera Manager will restart it automatically.

#### Prerequisites

To use work preserving recovery for the ResourceManager, you need to first enable High Availability for the ResourceManager. See [YARN \(MRv2\) ResourceManager High Availability](#) on page 237 for more information.

### Configuring Work Preserving Recovery Using Cloudera Manager

Use this procedure if you manage your CDH cluster using Cloudera Manager. Otherwise, see [Configuring Work Preserving Recovery Using the Command Line](#) on page 244.

If you use Cloudera Manager and you enable [YARN \(MRv2\) ResourceManager High Availability](#) on page 237, work preserving recovery is enabled by default for the ResourceManager. To disable the feature for the ResourceManager, change the value of `yarn.resourcemanager.work-preserving-recovery.enabled` to `false` in the `yarn-site.xml` using an advanced configuration snippet.

To enable the feature for a given NodeManager, edit the advanced configuration snippet for `yarn-site.xml` on that NodeManager, and set the value of `yarn.nodemanager.recovery.enabled` to `true`. For a given NodeManager, you can configure the directory on the local filesystem where state information is stored when work preserving recovery is enabled, by setting the value of `yarn.nodemanager.recovery.dir` to a local filesystem directory, using the same advanced configuration snippet. The default value is `${hadoop.tmp.dir}/yarn-nm-recovery`. This location usually points to the `/tmp` directory on the local filesystem. Because many operating systems do not preserve the contents of the `/tmp` directory across a reboot, Cloudera strongly recommends changing the location of `yarn.nodemanager.recovery.dir` to a different directory on the local filesystem. The [example](#) below uses `/home/cloudera/recovery`.

### Configuring Work Preserving Recovery Using the Command Line

- **Important:**
  - If you use Cloudera Manager, do not use these command-line instructions.
  - This information applies specifically to CDH 5.3.x. If you use an earlier version of CDH, see the documentation for that version located at [Cloudera Documentation](#).

After enabling ResourceManager High Availability, edit `yarn-site.xml` on the ResourceManager and all NodeManagers.

1. Set the value of `yarn.resourcemanager.work-preserving-recovery.enabled` to `true` to enable work preserving recovery for the ResourceManager, and set the value of `yarn.nodemanager.recovery.enabled` to `true` for the NodeManager.
2. For each NodeManager, configure the directory on the local filesystem where state information is stored when work preserving recovery is enabled, by setting the value of `yarn.nodemanager.recovery.dir` to a local filesystem directory. The default value is `${hadoop.tmp.dir}/yarn-nm-recovery`. This location usually points to the `/tmp` directory on the local filesystem. Because many operating systems do not preserve the contents of the `/tmp` directory across a reboot, Cloudera strongly recommends changing the location of `yarn.nodemanager.recovery.dir` to a different directory on the local filesystem. The [example](#) below uses `/home/cloudera/recovery`.
3. Configure a valid RPC address for the NodeManager, by setting `yarn.nodemanager.address` to an address with a specific port number (such as `0.0.0.0:45454`). Ephemeral ports (port 0, which is default) cannot be used for the NodeManager's RPC server, because this could cause the NodeManager to use different ports before and after a restart, preventing clients from connecting to the NodeManager after a restart. The NodeManager RPC address is also important for auxiliary services which run in a YARN cluster.

Auxiliary services should also be designed to support recoverability by reloading the previous state after a NodeManager restarts. An example auxiliary service, the ShuffleHandler service for MapReduce, follows the correct pattern for an auxiliary service which supports work preserving recovery of the NodeManager.

### Example Configuration for Work Preserving Recovery

The following example configuration can be used with a Cloudera Manager advanced configuration snippet or added to `yarn-site.xml` directly if you do not use Cloudera Manager. Adjust the configuration to suit your environment.

```
<property>
  <name>yarn.resourcemanager.work-preserving-recovery.enabled</name>
  <value>true</value>
  <description>Whether to enable work preserving recovery for the Resource
Manager</description>
</property>
<property>
  <name>yarn.nodemanager.recovery.enabled</name>
  <value>true</value>
  <description>Whether to enable work preserving recovery for the Node
Manager</description>
</property>
<property>
  <name>yarn.nodemanager.recovery.dir</name>
  <value>/home/cloudera/recovery</value>
  <description>The location for stored state on the Node Manager, if work preserving
recovery
  is enabled.</description>
</property>
<property>
  <name>yarn.nodemanager.address</name>
  <value>0.0.0.0:45454</value>
</property>
```

## MapReduce (MRv1) JobTracker High Availability

Follow the instructions in this section to configure high availability (HA) for JobTracker.

### Configuring MapReduce (MRv1) JobTracker High Availability Using Cloudera Manager

**Required Role:** Cluster Administrator Full Administrator

You can use Cloudera Manager to configure CDH 4.3 or later for JobTracker high availability (HA). Although it is possible to configure JobTracker HA with CDH 4.2, it is not recommended. Rolling restart, decommissioning of

## High Availability

TaskTrackers, and rolling upgrade of MapReduce from CDH 4.2 to CDH 4.3 are not supported when JobTracker HA is enabled.

Cloudera Manager supports automatic failover of the JobTracker. It does not provide a mechanism to manually force a failover through the Cloudera Manager user interface.

- **Important:** Enabling or disabling JobTracker HA will cause the previous monitoring history to become unavailable.

### Enabling JobTracker High Availability

The **Enable High Availability** workflow leads you through adding a second (standby) JobTracker:

1. Go to the MapReduce service.
2. Select **Actions** > **Enable High Availability**. A screen showing the hosts that are eligible to run a standby JobTracker displays. The host where the current JobTracker is running is not available as a choice.
3. Select the host where you want the Standby JobTracker to be installed, and click **Continue**.
4. Enter a directory location on the local filesystem for each JobTracker host. These directories will be used to store job configuration data.
  - You may enter more than one directory, though it is not required. The paths do not need to be the same on both JobTracker hosts.
  - If the directories you specify do not exist, they will be created with the appropriate permissions. If they already exist, they must be empty and have the appropriate permissions.
  - If the directories are not empty, Cloudera Manager will not delete the contents.
5. Optionally use the checkbox under Advanced Options to force initialize the ZooKeeper znode for auto-failover.
6. Click **Continue**. Cloudera Manager executes a set of commands that stop the MapReduce service, add a standby JobTracker and Failover controller, initialize the JobTracker high availability state in ZooKeeper, create the job status directory, restart MapReduce, and redeploy the relevant client configurations.

### Disabling JobTracker High Availability

1. Go to the MapReduce service.
2. Select **Actions** > **Disable High Availability**. A screen showing the hosts running the JobTrackers displays.
3. Select which JobTracker (host) you want to remain as the single JobTracker, and click **Continue**. Cloudera Manager executes a set of commands that stop the MapReduce service, remove the standby JobTracker and the Failover Controller, restart the MapReduce service, and redeploy client configurations.

### Configuring MapReduce (MRv1) JobTracker High Availability Using the Command Line

If you are running MRv1, you can configure the JobTracker to be highly available. You can configure either manual or automatic failover to a warm-standby JobTracker.

- **Note:**
  - As with [HDFS High Availability](#) on page 211, the JobTracker high availability feature is backward compatible; that is, if you do not want to enable JobTracker high availability, you can simply keep your existing configuration after updating your `hadoop-0.20-mapreduce`, `hadoop-0.20-mapreduce-jobtracker`, and `hadoop-0.20-mapreduce-tasktracker` packages, and start your services as before. You do not need to perform any of the actions described on this page.

To use the high availability feature, you must create a new configuration. This new configuration is designed such that all the hosts in the cluster can have the same configuration; you do not need to deploy different configuration files to different hosts depending on each host's role in the cluster.

In an HA setup, the `mapred.job.tracker` property is no longer a `host:port` string, but instead specifies a logical name to identify JobTracker instances in the cluster (active and standby). Each distinct JobTracker in the cluster has a different JobTracker ID. To support a single configuration file for all of the JobTrackers, the relevant configuration parameters are suffixed with the JobTracker logical name as well as the JobTracker ID.

The HA JobTracker is packaged separately from the original (non-HA) JobTracker.

- **Important:** You cannot run both HA and non-HA JobTrackers in the same cluster. Do not install the HA JobTracker unless you need a highly available JobTracker. If you install the HA JobTracker and later decide to revert to the non-HA JobTracker, you will need to uninstall the HA JobTracker and re-install the non-HA JobTracker.

JobTracker HA reuses the `mapred.job.tracker` parameter in `mapred-site.xml` to identify a JobTracker active-standby pair. In addition, you must enable the existing `mapred.jobtracker.restart.recover`, `mapred.job.tracker.persist.jobstatus.active`, and `mapred.job.tracker.persist.jobstatus.hours` parameters, as well as a number of new parameters, as discussed below.

Use the sections that follow to install, configure and test JobTracker HA.

### Replacing the non-HA JobTracker with the HA JobTracker

This section provides instructions for removing the non-HA JobTracker and installing the HA JobTracker.

- **Important:** The HA JobTracker cannot be installed on a node on which the non-HA JobTracker is installed, and vice versa. If the JobTracker is installed, uninstall it following the instructions below before installing the HA JobTracker. Uninstall the non-HA JobTracker whether or not you intend to install the HA JobTracker on the same node.

### Removing the non-HA JobTracker

You must remove the original (non-HA) JobTracker before you install and run the HA JobTracker. First, you need to stop the JobTracker and TaskTrackers.

#### To stop the JobTracker and TaskTrackers:

1. Stop the TaskTrackers: On each TaskTracker system:

```
$ sudo service hadoop-0.20-mapreduce-tasktracker stop
```

2. Stop the JobTracker: On the JobTracker system:

```
$ sudo service hadoop-0.20-mapreduce-jobtracker stop
```

3. Verify that the JobTracker and TaskTrackers have stopped:

```
$ ps -eaf | grep -i job
$ ps -eaf | grep -i task
```

#### To remove the JobTracker:

- On Red Hat-compatible systems:

```
$ sudo yum remove hadoop-0.20-mapreduce-jobtracker
```

- On SLES systems:

```
$ sudo zypper remove hadoop-0.20-mapreduce-jobtracker
```

- On Ubuntu systems:

```
sudo apt-get remove hadoop-0.20-mapreduce-jobtracker
```

### Installing the HA JobTracker

Use the following steps to install the HA JobTracker package, and optionally the ZooKeeper failover controller package (needed for automatic failover).

#### Step 1: Install the HA JobTracker package on two separate nodes

##### On each JobTracker node:

- On Red Hat-compatible systems:

```
$ sudo yum install hadoop-0.20-mapreduce-jobtrackerha
```

- On SLES systems:

```
$ sudo zypper install hadoop-0.20-mapreduce-jobtrackerha
```

- On Ubuntu systems:

```
sudo apt-get install hadoop-0.20-mapreduce-jobtrackerha
```

#### Step 2: (Optionally) install the failover controller package

If you intend to enable automatic failover, you need to install the failover controller package.

- **Note:** The [instructions for automatic failover](#) assume that you have set up a ZooKeeper cluster running on three or more nodes, and have verified its correct operation by connecting using the ZooKeeper command-line interface (CLI). See the [ZooKeeper documentation](#) for instructions on how to set up a ZooKeeper ensemble.

Install the failover controller package as follows:

##### On each JobTracker node:

- On Red Hat-compatible systems:

```
$ sudo yum install hadoop-0.20-mapreduce-zkfc
```

- On SLES systems:

```
$ sudo zypper install hadoop-0.20-mapreduce-zkfc
```

- On Ubuntu systems:

```
sudo apt-get install hadoop-0.20-mapreduce-zkfc
```

### Configuring and Deploying Manual Failover

Proceed as follows to configure manual failover:

1. [Configure the JobTrackers, TaskTrackers, and Clients](#)
2. [Start the JobTrackers](#)
3. [Activate a JobTracker](#)
4. [Verify that failover is working](#)



## Step 1: Configure the JobTrackers, TaskTrackers, and Clients

## Changes to existing configuration parameters

Property name	Default	Used on	Description
<code>mapred.job.tracker</code>	<code>local</code>	JobTracker, TaskTracker, client	In an HA setup, the logical name of the JobTracker active-standby pair. In a non-HA setup <code>mapred.job.tracker</code> is a <code>host:port</code> string specifying the JobTracker's RPC address, but in an HA configuration the logical name <i>must not</i> include a port number.
<code>mapred.jobtracker.restart.recover</code>	<code>false</code>	JobTracker	Whether to recover jobs that were running in the most recent active JobTracker. Must be set to <code>true</code> for JobTracker HA.
<code>mapred.job.tracker.persist.jobstatus.active</code>	<code>false</code>	JobTracker	Whether to make job status persistent in HDFS. Must be set to <code>true</code> for JobTracker HA.
<code>mapred.job.tracker.persist.jobstatus.hours</code>	<code>0</code>	JobTracker	The number of hours job status information is retained in HDFS. Must be greater than zero for JobTracker HA.
<code>mapred.job.tracker.persist.jobstatus.dir</code>	<code>/jobtracker/jobstatus</code>	JobTracker	The HDFS directory in which job status information is kept persistently. The directory must exist and be owned by the <code>mapred</code> user.

## New configuration parameters

Property name	Default	Used on	Description
<code>mapred.jobtrackers.&lt;name&gt;</code>	None	JobTracker, TaskTracker, client	A comma-separated pair of IDs for the active and standby JobTrackers. The <code>&lt;name&gt;</code> is the value of <code>mapred.job.tracker</code> .
<code>mapred.jobtracker.rpc-address.&lt;name&gt;.&lt;id&gt;</code>	None	JobTracker, TaskTracker, client	The RPC address of an individual JobTracker. <code>&lt;name&gt;</code> refers to the value of <code>mapred.job.tracker</code> ; <code>&lt;id&gt;</code> refers to one or other of the

Property name	Default	Used on	Description
			values in <code>mapred.jobtrackers.&lt;name&gt;.</code>
<code>mapred.job.tracker.http.address.&lt;name&gt;.&lt;id&gt;</code>	None	JobTracker, TaskTracker	The HTTP address of an individual JobTracker. (In a non-HA setup <code>mapred.job.tracker.http.address</code> (with no suffix) is the JobTracker's HTTP address.)
<code>mapred.ha.jobtracker.rpc-address.&lt;name&gt;.&lt;id&gt;</code>	None	JobTracker, failover controller	The RPC address of the HA service protocol for the JobTracker. The JobTracker listens on a separate port for HA operations which is why this property exists in addition to <code>mapred.jobtracker.rpc-address.&lt;name&gt;.&lt;id&gt;</code>
<code>mapred.ha.jobtracker.http-direct-address.&lt;name&gt;.&lt;id&gt;</code>	None	JobTracker	The HTTP address of an individual JobTracker that should be used for HTTP redirects. The standby JobTracker will redirect all web traffic to the active, and will use this property to discover the URL to use for redirects. A property separate from <code>mapred.job.tracker.http.address.&lt;name&gt;.&lt;id&gt;</code> is needed since the latter may be a wildcard bind address, such as <code>0.0.0.0:50030</code> , which is not suitable for making requests. Note also that <code>mapred.job.tracker.http-address.&lt;name&gt;.&lt;id&gt;</code> is the HTTP redirect address for the JobTracker with ID <code>&lt;id&gt;</code> for the pair with the logical name <code>&lt;name&gt;</code> - that is, the address that should be used when that JobTracker is active, and <i>not</i> the address that should be redirected to when that JobTracker is the standby.
<code>mapred.ha.jobtracker.id</code>	None	JobTracker	The identity of this JobTracker instance. Note that this is optional since each JobTracker can infer its ID from the matching address in one of the

Property name	Default	Used on	Description
			<code>mapred.jobtracker.rpc-address.&lt;name&gt;.&lt;id&gt;</code> properties. It is provided for testing purposes.
<code>mapred.client.failover.proxy.provider.&lt;name&gt;</code>	None	TaskTracker, client	The failover provider class. The only class available is <code>org.apache.hadoop.mapred.ConfiguredFailoverProxyProvider</code> .
<code>mapred.client.failover.max.attempts</code>	15	TaskTracker, client	The maximum number of times to try to fail over.
<code>mapred.client.failover.sleep.base.millis</code>	500	TaskTracker, client	The time to wait before the first failover.
<code>mapred.client.failover.sleep.max.millis</code>	1500	TaskTracker, client	The maximum amount of time to wait between failovers (for exponential backoff).
<code>mapred.client.failover.connection.retries</code>	0	TaskTracker, client	The maximum number of times to retry between failovers.
<code>mapred.client.failover.connection.retries.on.timeouts</code>	0	TaskTracker, client	The maximum number of times to retry on timeouts between failovers.
<code>mapred.ha.fencing.methods</code>	None	failover controller	<p>A list of scripts or Java classes that will be used to fence the active JobTracker during failover.</p> <p>Only one JobTracker should be active at any given time, but you can simply configure <code>mapred.ha.fencing.methods</code> as <code>shell(/bin/true)</code> since the JobTrackers fence themselves, and split-brain is avoided by the old active JobTracker shutting itself down if another JobTracker takes over.</p>

Make changes and additions similar to the following to `mapred-site.xml` on each node.

- **Note:** It is simplest to configure all the parameters on all nodes, even though not all of the parameters will be used on any given node. This also makes for robustness if you later change the roles of the nodes in your cluster.

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
```

```

<!-- Put site-specific property overrides in this file. -->

<configuration>

  <property>
    <name>mapred.job.tracker</name>
    <value>logicaljt</value>
    <!-- host:port string is replaced with a logical name -->
  </property>

  <property>
    <name>mapred.jobtrackers.logicaljt</name>
    <value>jt1,jt2</value>
    <description>Comma-separated list of JobTracker IDs.</description>
  </property>

  <property>
    <name>mapred.jobtracker.rpc-address.logicaljt.jt1</name>
    <!-- RPC address for jt1 -->
    <value>myjt1.myco.com:8021</value>
  </property>

  <property>
    <name>mapred.jobtracker.rpc-address.logicaljt.jt2</name>
    <!-- RPC address for jt2 -->
    <value>myjt2.myco.com:8022</value>
  </property>

  <property>
    <name>mapred.job.tracker.http.address.logicaljt.jt1</name>
    <!-- HTTP bind address for jt1 -->
    <value>0.0.0.0:50030</value>
  </property>

  <property>
    <name>mapred.job.tracker.http.address.logicaljt.jt2</name>
    <!-- HTTP bind address for jt2 -->
    <value>0.0.0.0:50031</value>
  </property>

  <property>
    <name>mapred.ha.jobtracker.rpc-address.logicaljt.jt1</name>
    <!-- RPC address for jt1 HA daemon -->
    <value>myjt1.myco.com:8023</value>
  </property>

  <property>
    <name>mapred.ha.jobtracker.rpc-address.logicaljt.jt2</name>
    <!-- RPC address for jt2 HA daemon -->
    <value>myjt2.myco.com:8024</value>
  </property>

  <property>
    <name>mapred.ha.jobtracker.http-redirect-address.logicaljt.jt1</name>
    <!-- HTTP redirect address for jt1 -->
    <value>myjt1.myco.com:50030</value>
  </property>

  <property>
    <name>mapred.ha.jobtracker.http-redirect-address.logicaljt.jt2</name>
    <!-- HTTP redirect address for jt2 -->
    <value>myjt2.myco.com:50031</value>
  </property>

  <property>
    <name>mapred.jobtracker.restart.recover</name>
    <value>true</value>
  </property>

  <property>
    <name>mapred.job.tracker.persist.jobstatus.active</name>
    <value>true</value>
  </property>

```

```

    </property>
<property>
  <name>mapred.job.tracker.persist.jobstatus.hours</name>
  <value>1</value>
</property>

<property>
  <name>mapred.job.tracker.persist.jobstatus.dir</name>
  <value>/jobtracker/jobsInfo</value>
</property>

<property>
  <name>mapred.client.failover.proxy.provider.logicaljt</name>
  <value>org.apache.hadoop.mapred.ConfiguredFailoverProxyProvider</value>
</property>

<property>
  <name>mapred.client.failover.max.attempts</name>
  <value>15</value>
</property>

<property>
  <name>mapred.client.failover.sleep.base.millis</name>
  <value>500</value>
</property>

<property>
  <name>mapred.client.failover.sleep.max.millis</name>
  <value>1500</value>
</property>

<property>
  <name>mapred.client.failover.connection.retries</name>
  <value>0</value>
</property>

<property>
  <name>mapred.client.failover.connection.retries.on.timeouts</name>
  <value>0</value>
</property>
<property>
  <name>mapred.ha.fencing.methods</name>
  <value>shell(/bin/true)</value>
</property>
</configuration>

```

■ **Note:**

In pseudo-distributed mode you need to specify `mapred.ha.jobtracker.id` for each JobTracker, so that the JobTracker knows its identity.

But in a fully-distributed setup, where the JobTrackers run on different nodes, there is no need to set `mapred.ha.jobtracker.id`, since the JobTracker can infer the ID from the matching address in one of the `mapred.jobtracker.rpc-address.<name>.<id>` properties.

## Step 2: Start the JobTracker daemons

To start the daemons, run the following command on each JobTracker node:

```
$ sudo service hadoop-0.20-mapreduce-jobtrackerha start
```

### Step 3: Activate a JobTracker

■ **Note:**

- You must be the `mapred` user to use `mrhaadmin` commands.
- If Kerberos is enabled, do not use `sudo -u mapred` when using the `hadoop mrhaadmin` command. Instead, you must log in with the `mapred` Kerberos credentials (the short name must be `mapred`). See [Configuring Hadoop Security in CDH 5](#) for more information.

Unless automatic failover is configured, both JobTrackers will be in a standby state after the `jobtrackerha` daemons start up.

If Kerberos is not enabled, use the following commands:

**To find out what state each JobTracker is in:**

```
$ sudo -u mapred hadoop mrhaadmin -getServiceState <id>
```

where `<id>` is one of the values you [configured](#) in the `mapred.jobtrackers.<name>` property – `jt1` or `jt2` in our sample `mapred-site.xml` files.

**To transition one of the JobTrackers to active and then verify that it is active:**

```
$ sudo -u mapred hadoop mrhaadmin -transitionToActive <id>
$ sudo -u mapred hadoop mrhaadmin -getServiceState <id>
```

where `<id>` is one of the values you [configured](#) in the `mapred.jobtrackers.<name>` property – `jt1` or `jt2` in our sample `mapred-site.xml` files.

With Kerberos enabled, log in as the `mapred` user and use the following commands:

**To log in as the `mapred` user and kinit:**

```
$ sudo su - mapred
$ kinit -kt mapred.keytab mapred/<fully.qualified.domain.name>
```

**To find out what state each JobTracker is in:**

```
$ hadoop mrhaadmin -getServiceState <id>
```

where `<id>` is one of the values you [configured](#) in the `mapred.jobtrackers.<name>` property – `jt1` or `jt2` in our sample `mapred-site.xml` files.

**To transition one of the JobTrackers to active and then verify that it is active:**

```
$ hadoop mrhaadmin -transitionToActive <id>
$ hadoop mrhaadmin -getServiceState <id>
```

where `<id>` is one of the values you [configured](#) in the `mapred.jobtrackers.<name>` property – `jt1` or `jt2` in our sample `mapred-site.xml` files.

### Step 4: Verify that failover is working

Use the following commands, depending whether or not Kerberos is enabled.

If Kerberos is not enabled, use the following commands:

**To cause a failover from the currently active to the currently inactive JobTracker:**

```
$ sudo -u mapred hadoop mrhaadmin -failover <id_of_active_JobTracker>
<id_of_inactive_JobTracker>
```

For example, if jt1 is currently active:

```
$ sudo -u mapred hadoop mrhaadmin -failover jt1 jt2
```

**To verify the failover:**

```
$ sudo -u mapred hadoop mrhaadmin -getServiceState <id>
```

For example, if jt2 should now be active:

```
$ sudo -u mapred hadoop mrhaadmin -getServiceState jt2
```

With Kerberos enabled, use the following commands:

**To log in as the mapred user and kinit:**

```
$ sudo su - mapred
$ kinit -kt mapred.keytab mapred/<fully.qualified.domain.name>
```

**To cause a failover from the currently active to the currently inactive JobTracker:**

```
$ hadoop mrhaadmin -failover <id_of_active_JobTracker> <id_of_inactive_JobTracker>
```

For example, if jt1 is currently active:

```
$ hadoop mrhaadmin -failover jt1 jt2
```

**To verify the failover:**

```
$ hadoop mrhaadmin -getServiceState <id>
```

For example, if jt2 should now be active:

```
$ hadoop mrhaadmin -getServiceState jt2
```

## Configuring and Deploying Automatic Failover

To configure automatic failover, proceed as follows:

1. [Configure a ZooKeeper ensemble](#) (if necessary)
2. [Configure parameters for manual failover](#)
3. [Configure failover controller parameters](#)
4. [Initialize the HA state in ZooKeeper](#)
5. [Enable automatic failover](#)
6. [Verify automatic failover](#)

### Step 1: Configure a ZooKeeper ensemble (if necessary)

To support automatic failover you need to set up a ZooKeeper ensemble running on three or more nodes, and verify its correct operation by connecting using the ZooKeeper command-line interface (CLI). See the [ZooKeeper documentation](#) for instructions on how to set up a ZooKeeper ensemble.

- **Note:** If you are already using a ZooKeeper ensemble for [automatic failover](#), use the same ensemble for automatic JobTracker failover.

Step 2: Configure the parameters for manual failover

See the instructions for configuring the TaskTrackers and JobTrackers under [Configuring and Deploying Manual Failover](#).

Step 3: Configure failover controller parameters

Use the following additional parameters to configure a failover controller for each JobTracker. The failover controller daemons run on the JobTracker nodes.

New configuration parameters

Property name	Default	Configure on	Description
<del>mapred.ha.automatic-failover.enabled</del> mapred.ha.automatic-failover.enabled	false	failover controller	Set to <code>true</code> to enable automatic failover.
mapred.ha.zkfc.port	8019	failover controller	The ZooKeeper failover controller port.
ha.zookeeper.quorum	None	failover controller	The ZooKeeper quorum (ensemble) to use for MRZKFailoverController.

Add the following configuration information to `mapred-site.xml`:

```
<property>
  <name>mapred.ha.automatic-failover.enabled</name>
  <value>true</value>
</property>

<property>
  <name>mapred.ha.zkfc.port</name>
  <value>8018</value>
  <!-- Pick a different port for each failover controller when running one machine -->
</property>
```

Add an entry similar to the following to `core-site.xml`:

```
<property>
  <name>ha.zookeeper.quorum</name>
  <value>zk1.example.com:2181,zk2.example.com:2181,zk3.example.com:2181 </value>
  <!-- ZK ensemble addresses -->
</property>
```

- **Note:** If you have already [configured automatic failover for HDFS](#), this property is already properly configured; you use the same ZooKeeper ensemble for HDFS and JobTracker HA.

Step 4: Initialize the HA State in ZooKeeper

After you have configured the failover controllers, the next step is to initialize the required state in ZooKeeper. You can do so by running one of the following commands from one of the JobTracker nodes.



- **Note:** The ZooKeeper ensemble must be running when you use this command; otherwise it will not work properly.

```
$ sudo service hadoop-0.20-mapreduce-zkfc init
```

or

```
$ sudo -u mapred hadoop mrzkfc -formatZK
```

This will create a `znode` in ZooKeeper in which the automatic failover system stores its data.

- **Note:** If you are running a secure cluster, see also [Securing access to ZooKeeper](#).

### Step 5: Enable automatic failover

To enable automatic failover once you have completed the configuration steps, you need only start the `jobtrackerha` and `zkfc` daemons.

To start the daemons, run the following commands on each JobTracker node:

```
$ sudo service hadoop-0.20-mapreduce-zkfc start
$ sudo service hadoop-0.20-mapreduce-jobtrackerha start
```

One of the JobTrackers will automatically transition to active.

### Step 6: Verify automatic failover

After enabling automatic failover, you should test its operation. To do so, first locate the active JobTracker. To find out what state each JobTracker is in, use the following command:

```
$ sudo -u mapred hadoop mrhaadmin -getServiceState <id>
```

where `<id>` is one of the values you [configured](#) in the `mapred.jobtrackers.<name>` property – `jt1` or `jt2` in our sample `mapred-site.xml` files.

- **Note:** You must be the `mapred` user to use `mrhaadmin` commands.

Once you have located your active JobTracker, you can cause a failure on that node. For example, you can use `kill -9 <pid of JobTracker>` to simulate a JVM crash. Or you can power-cycle the machine or its network interface to simulate different kinds of outages. After you trigger the outage you want to test, the other JobTracker should automatically become active within several seconds. The amount of time required to detect a failure and trigger a failover depends on the configuration of `ha.zookeeper.session-timeout.ms`, but defaults to 5 seconds.

If the test does not succeed, you may have a misconfiguration. Check the logs for the `zkfc` and `jobtrackerha` daemons to diagnose the problem.

## Usage Notes

### Using the JobTracker Web UI

To use the JobTracker Web UI, use the HTTP address of either JobTracker (that is, the value of `mapred.job.tracker.http.address.<name>.<id>` for either the active or the standby JobTracker). Note the following:

- If you use the URL of the standby JobTracker, you will be redirected to the active JobTracker.
- If you use the URL of a JobTracker that is down, you *will not* be redirected – you will simply get a "Not Found" error from your browser.

### Turning off Job Recovery

After a failover, the newly active JobTracker by default restarts all jobs that were running when the failover occurred. For Sqoop 1 and HBase jobs, this is undesirable because they are not **idempotent** (that is, they do not behave the same way on repeated execution). For these jobs you should consider setting `mapred.job.restart.recover` to `false` in the job configuration (`JobConf`).

## High Availability for Other CDH Components

This section provides information on high availability for CDH components independently of HDFS. See also [Configuring Other CDH Components to Use HDFS HA](#) on page 231.

For details about HA for Impala, see [Using Impala through a Proxy for High Availability](#).

### HBase High Availability

Most aspects of HBase are highly available in a standard configuration. A cluster typically consists of one Master and three or more RegionServers, with data stored in HDFS. To ensure that every component is highly available, configure one or more backup Masters. The backup Masters run on other hosts than the active Master.

#### Enabling HBase High Availability Using Cloudera Manager

1. Go to the HBase service.
2. Follow the process for [adding a role instance](#) and add a backup Master to a different host than the one on which the active Master is running.

#### Enabling HBase High Availability Using the Command Line

To configure backup Masters, create a new file in the `conf/` directory which will be distributed across your cluster, called `backup-masters`. For each backup Master you wish to start, add a new line with the hostname where the Master should be started. Each host that will run a Master needs to have all of the configuration files available. In general, it is a good practice to distribute the entire `conf/` directory across all cluster nodes.

After saving and distributing the file, restart your cluster for the changes to take effect. When the master starts the backup Masters, messages are logged. In addition, you can verify that an `HMaster` process is listed in the output of the `jps` command on the nodes where the backup Master should be running.

```
$ jps
15930 HRegionServer
16194 Jps
15838 HQuorumPeer
16010 HMaster
```

To stop a backup Master without stopping the entire cluster, first find its process ID using the `jps` command, then issue the `kill` command against its process ID.

```
$ kill 16010
```

### Hive Metastore High Availability

You can enable Hive metastore high availability (HA), so that your cluster is resilient to failures due to a metastore that becomes unavailable. Each metastore is independent; they do not use a quorum.

#### Prerequisites


- Cloudera recommends that each instance of the metastore runs on a separate cluster host, to maximize high availability.
- Hive metastore HA requires a database that is also highly available, such as MySQL with replication. Refer to the documentation for your database of choice to configure it correctly.

## Limitations

Sentry HDFS synchronization does not support Hive metastore HA.

## Enabling Hive Metastore High Availability Using Cloudera Manager

Required Role: **Configurator** **Cluster Administrator** **Full Administrator**

1. Go to the Hive service.
2. Click the **Configuration** tab.
3. Click the **Advanced** category.
4. If you use a secure cluster, enable the Hive token store by setting the value of the **Hive Metastore Delegation Token Store** property to `org.apache.hadoop.hive.thrift.DBTokenStore`.
5. Click **Save Changes** to commit the changes.
6. Click the **Instances** tab.
7. Click **Add Role Instances**.
8. Click the text field under **Hive Metastore Server**.
9. Select a host on which to run the additional metastore and click **OK**.
10. Click **Continue** and click **Finish**.
11. Check the checkbox next to the new **Hive Metastore Server** role.
12. Select **Actions for Selected** > **Start**, and click **Start** to confirm.
13. Click  to display the stale configurations page.
14. Click **Restart Cluster** and click **Restart Now**.
15. Click **Finish** after the cluster finishes restarting.

## Enabling Hive Metastore High Availability Using the Command Line

To configure the Hive metastore for high availability, you configure each metastore to store its state in a replicated database, then provide the metastore clients with a list of URIs where metastores are available. The client starts with the first URI in the list. If it does not get a response, it randomly picks another URI in the list and attempts to connect. This continues until the client receives a response.

### Important:

- If you use Cloudera Manager, do not use these command-line instructions.
- This information applies specifically to CDH 5.3.x . If you use an earlier version of CDH, see the documentation for that version located at [Cloudera Documentation](#).

1. Configure Hive on each of the cluster hosts where you want to run a metastore, following the instructions at [Configuring the Hive Metastore](#).
2. On the server where the master metastore instance runs, edit the `/etc/hive/conf.server/hive-site.xml` file, setting the `hive.metastore.uris` property's value to a list of URIs where a Hive metastore is available for failover.

```
<property>
  <name>hive.metastore.uris</name>

  <value>thrift://metastore1.example.com,thrift://metastore2.example.com,thrift://metastore3.example.com</value>

  <description> URI for client to contact metastore server </description>
</property>
```

## High Availability

3. If you use a secure cluster, enable the Hive token store by configuring the value of the `hive.cluster.delegation.token.store.class` property to `org.apache.hadoop.hive.thrift.DBTokenStore`.

```
<property>
  <name>hive.cluster.delegation.token.store.class</name>
  <value>org.apache.hadoop.hive.thrift.DBTokenStore</value>
</property>
```

4. Save your changes and restart each Hive instance.
5. Connect to each metastore and update it to use a nameservice instead of a NameNode, as a requirement for high availability.
  - a. From the command-line, as the Hive user, retrieve the list of URIs representing the filesystem roots:  

```
hive --service metatool -listFSRoot
```
  - b. Run the following command with the `--dry-run` option, to be sure that the nameservice is available and configured correctly. This will not change your configuration.  

```
hive --service metatool -updateLocation nameservice-uri namenode-uri --dryRun
```
  - c. Run the same command again without the `--dry-run` option to direct the metastore to use the nameservice instead of a NameNode.  

```
hive --service metatool -updateLocation nameservice-uri namenode-uri
```
6. Test your configuration by stopping your main metastore instance, and then attempting to connect to one of the other metastores from a client. The following is an example of doing this on a RHEL or Fedora system. The example first stops the local metastore, then connects to the metastore on the host `metastore2.example.com` and runs the `SHOW TABLES` command.

```
$ sudo service hive-metastore stop
$ /usr/lib/hive/bin/beeline
beeline> !connect jdbc:hive2://metastore2.example.com:10000 username password
org.apache.hive.jdbc.HiveDriver
0: jdbc:hive2://localhost:10000> SHOW TABLES;
show tables;
+-----+
| tab_name |
+-----+
+-----+
No rows selected (0.238 seconds)
0: jdbc:hive2://localhost:10000>
```

7. Restart the local metastore when you have finished testing.

```
$ sudo service hive-metastore stop
```

## Llama High Availability

Llama high availability (HA) uses an active-standby architecture, in which the active Llama is automatically elected using the ZooKeeper-based `ActiveStandbyElector`. The active Llama accepts RPC Thrift connections and communicates with YARN. The standby Llama monitors the leader information in ZooKeeper, but doesn't accept RPC Thrift connections.

### Fencing

Only one of the Llamas should be active to ensure the resources are not partitioned. Llama uses ZooKeeper access control lists (ACLs) to claim exclusive ownership of the cluster when transitioning to active, and monitors this ownership periodically. If another Llama takes over, the first one realizes it within this period.

## Reclaiming Cluster Resources

To claim resources from YARN, Llama spawns YARN applications and runs unmanaged ApplicationMasters. When a Llama goes down, the resources allocated to all the YARN applications spawned by it are not reclaimed until YARN times out those applications (the default timeout is 10 minutes). On Llama failure, these resources are reclaimed by means of a Llama that kills any YARN applications spawned by this pair of Llamas.

## Enabling Llama High Availability Using Cloudera Manager

You can enable Llama high availability when you [enable integrated resource management](#). If you chose to create a single Llama instance at that time, follow these steps to enable Llama high availability:

1. Go to the Impala service.
2. Select **Actions > Enable High Availability**.
3. Click the **Impala Llama ApplicationMaster Hosts** field to display a dialog for choosing Llama hosts.

The following shortcuts for specifying hostname patterns are supported:

- Range of hostnames (without the domain portion)

Range Definition	Matching Hosts
10.1.1.[1-4]	10.1.1.1, 10.1.1.2, 10.1.1.3, 10.1.1.4
host[1-3].company.com	host1.company.com, host2.company.com, host3.company.com
host[07-10].company.com	host07.company.com, host08.company.com, host09.company.com, host10.company.com

- IP addresses
- Rack name

4. Specify or select one or more hosts and click **OK**.
5. Click **Continue**.
6. Click **Continue**. A progress screen displays with a summary of the wizard actions.
7. Click **Finish**.

## Disabling Llama High Availability Using Cloudera Manager

You can disable Llama high availability when you [disable integrated resource management](#). If you choose not to disable integrated resource management, follow these steps to disable Llama high availability:

1. Go to the Impala service.
2. Select **Actions > Disable High Availability**.
3. Choose the host on which Llama runs after high availability is disabled.
4. Click **Continue**. A progress screen displays with a summary of the wizard actions.
5. Click **Finish**.

## Enabling Llama High Availability Using the Command Line

1. Configure Llama HA by modifying the following configuration properties in `/etc/llama/conf/llama-site.xml`. There is no need for any additional daemons.

Property	Description	Default	Recommended
<code>llama.am.cluster.id</code>	Cluster ID of the Llama pair, used to differentiate between different Llamas	<code>llama</code>	[cluster-specific]
<code>llama.am.ha.enabled*</code>	Whether to enable Llama HA	<code>false</code>	<code>true</code>

Property	Description	Default	Recommended
<code>llama.am.ha.zk-quorum*</code>	ZooKeeper quorum to use for leader election and fencing		[cluster-specific]
<code>llama.am.ha.zk-base</code>	Base znode for leader election and fencing data	<code>/llama</code>	[cluster-specific]
<code>llama.am.ha.zk-timeout-ms</code>	The session timeout, in milliseconds, for connections to ZooKeeper quorum	10000	10000
<code>llama.am.ha.zk-acl</code>	ACLs to control access to ZooKeeper	<code>world:anyone:rwcd</code>	[cluster-specific]
<code>llama.am.ha.zk-auth</code>	Authentication information to go with the ACLs		[cluster-acl-specific]

\*Required configurations

2. Configure the [Impala daemon to use the HA Llama](#).

## Oozie High Availability

In CDH 5, you can configure multiple active Oozie servers against the same database, providing high availability for Oozie. This is supported in both MRv1 or MRv2 (YARN).

### Requirements

The requirements for Oozie high availability are:

- An external database that supports multiple concurrent connections. The default Derby database does not support multiple concurrent connections. In addition, the database should be configured for HA (for example Oracle RAC, MySQL Cluster). If the database is not HA and fails all Oozie servers will stop working. HA will still work with a non-HA database, but then the database then becomes the single point of failure.
- On all the hosts where Oozie servers are going to run, the JDBC JAR should be placed in `/var/lib/oozie/` or in the location referenced by the environment variables `CLOUDERA_MYSQL_CONNECTOR_JAR` or `CLOUDERA_ORACLE_CONNECTOR_JAR` if using MySQL or Oracle respectively.
- ZooKeeper, which is used for distributed locks to coordinate the Oozie servers accessing the database at the same time and service discovery so that the Oozie servers can locate each other for log aggregation.
- A load balancer that
  - A load balancer (preferably with HA support, for example [HAProxy](#)), Virtual IP, or Round-Robin DNS, to provide a single entry-point for users so they don't have to choose between, or even be aware of, multiple Oozie servers and for callbacks from the ApplicationMaster or JobTracker
  - Receives callbacks from JobTracker when a job is done. Callbacks are best-effort and used as "hints", so eventually, default is  $\leq 10\text{min}$ , the other Oozie servers would go and contact the JobTracker regardless of whether or not the callback went through and nothing would be lost or stuck. The load balancer should be HA as well. The load balancer should be configured for round robin and not take into account the actual load on any of the Oozie servers.

For information on setting up SSL communication with Oozie HA enabled, see [Additional Considerations when Configuring SSL for Oozie HA](#).

## Enabling Oozie High Availability Using Cloudera Manager

Required Role: **Full Administrator**

- **Important:** Enabling or disabling HA will cause the previous monitoring history to become unavailable.

### Enabling Oozie High Availability

1. Ensure that the [requirements](#) are satisfied.
2. In the Cloudera Manager Admin Console, go to the Oozie service.
3. Select **Actions** > **Enable High Availability**. A screen showing the hosts that are eligible to run an additional Oozie server displays. The host where the current Oozie server is running is not available as a choice.
4. Select the host where you want the additional Oozie server to be installed, and click **Continue**.
5. Specify the host and port of the Oozie load balancer, and click **Continue**. Cloudera Manager executes a set of commands that stops Oozie servers, add another Oozie server, initializes the Oozie server High Availability state in ZooKeeper, configures Hue to reference the Oozie load balancer, and restarts the Oozie servers and dependent services.

### Disabling Oozie High Availability

1. In the Cloudera Manager Admin Console, go to the Oozie service.
2. Select **Actions** > **Disable High Availability**. A screen showing the hosts running the Oozie servers displays.
3. Select which Oozie server (host) you want to remain as the single Oozie server, and click **Continue**. Cloudera Manager executes a set of commands that stop the Oozie service, removes the additional Oozie servers, configures Hue to reference the Oozie service, and restarts the Oozie service and dependent services.

## Enabling Oozie High Availability Using the Command Line

For more information, and installation and configuration instructions for configuring Oozie HA using the command line, see <http://archive.cloudera.com/cdh5/cdh/5/oozie>.

## Search High Availability

Mission critical, large-scale online production systems need to make progress without downtime despite some issues. Cloudera Search provides two routes to configurable, highly available, and fault-tolerant data ingestion:

- Near Real Time (NRT) ingestion using the Flume Solr Sink
- MapReduce based batch ingestion using the MapReduceIndexerTool

### Production versus Test Mode

Some exceptions are generally transient, in which case the corresponding task can simply be retried. For example, network connection errors or timeouts are recoverable exceptions. Conversely, tasks associated with an unrecoverable exception cannot simply be retried. Corrupt or malformed parser input data, parser bugs, and errors related to unknown Solr schema fields produce unrecoverable exceptions.

Different modes determine how Cloudera Search responds to different types of exceptions.

- **Configuration parameter `isProductionMode=false`** (Non-production mode or test mode): Default configuration. Cloudera Search throws exceptions to quickly reveal failures, providing better debugging diagnostics to the user.
- **Configuration parameter `isProductionMode=true`** (Production mode): Cloudera Search logs and ignores unrecoverable exceptions, enabling mission-critical large-scale online production systems to make progress without downtime, despite some issues.

- **Note:** Categorizing exceptions as recoverable or unrecoverable addresses most cases, though it is possible that an unrecoverable exception could be accidentally misclassified as recoverable. Cloudera provides the `isIgnoringRecoverableExceptions` configuration parameter to address such a case. In a production environment, if an unrecoverable exception is discovered that is classified as recoverable, change `isIgnoringRecoverableExceptions` to `true`. Doing so allows systems to make progress and avoid retrying an event forever. This configuration flag should only be enabled if a misclassification bug has been identified. Please report such bugs to Cloudera.

If Cloudera Search throws an exception according to the rules described above, the caller, meaning Flume Solr Sink and MapReduceIndexerTool, can catch the exception and retry the task if it meets the criteria for such retries.

### Near Real Time Indexing with the Flume Solr Sink

The Flume Solr Sink uses the settings established by the `isProductionMode` and `isIgnoringRecoverableExceptions` parameters. If a SolrSink does nonetheless receive an exception, the SolrSink rolls the transaction back and pauses. This causes the Flume channel, which is essentially a queue, to redeliver the transaction's events to the SolrSink approximately five seconds later. This redelivering of the transaction event retries the ingest to Solr. This process of rolling back, backing off, and retrying continues until ingestion eventually succeeds.

Here is a corresponding example Flume configuration file `flume.conf`:

```
agent.sinks.solrSink.isProductionMode = true
agent.sinks.solrSink.isIgnoringRecoverableExceptions = true
```

In addition, Flume SolrSink automatically attempts to load balance and failover among the hosts of a SolrCloud before it considers the transaction rollback and retry. Load balancing and failover is done with the help of ZooKeeper, which itself can be configured to be highly available.

Further, Cloudera Manager can configure Flume so it automatically restarts if its process crashes.

To tolerate extended periods of Solr downtime, you can configure Flume to use a high-performance transactional persistent queue in the form of a [FileChannel](#). A FileChannel can use any number of local disk drives to buffer significant amounts of data. For example, you might buffer many terabytes of events corresponding to a week of data. Further, using the [optional replicating channels](#) Flume feature, you can configure Flume to replicate the same data both into HDFS as well as into Solr. Doing so ensures that if the Flume SolrSink channel runs out of disk space, data delivery is still delivered to HDFS, and this data can later be ingested from HDFS into Solr using MapReduce.

Many machines with many Flume Solr Sinks and FileChannels can be used in a failover and load balancing configuration to improve high availability and scalability. Flume SolrSink servers can be either co-located with live Solr servers serving end user queries, or Flume SolrSink servers can be deployed on separate industry standard hardware for improved scalability and reliability. By spreading indexing load across a large number of Flume SolrSink servers you can improve scalability. Indexing load can be replicated across multiple Flume SolrSink servers for high availability, for example using Flume features such as [Load balancing Sink Processor](#).

### Batch Indexing with MapReduceIndexerTool

The Mappers and Reducers of the MapReduceIndexerTool follow the settings established by the `isProductionMode` and `isIgnoringRecoverableExceptions` parameters. However, if a Mapper or Reducer of the MapReduceIndexerTool does receive an exception, it does not retry at all. Instead it lets the MapReduce task fail and relies on the Hadoop Job Tracker to retry failed MapReduce task attempts several times according to standard Hadoop semantics. Cloudera Manager can configure the Hadoop Job Tracker to be highly available. On MapReduceIndexerTool startup, all data in the output directory is deleted if that output directory already exists. To retry an entire job that has failed, rerun the program using the same arguments.



For example:

```
hadoop ... MapReduceIndexerTool ... -D isProductionMode=true -D  
isIgnoringRecoverableExceptions=true ...
```

# Backup and Disaster Recovery

This guide describes the Cloudera Manager backup and disaster recovery (BDR) features, which provide an integrated, easy-to-use solution for enabling data protection in the Hadoop platform.

■ **Important:** This feature is available only with a Cloudera Enterprise license.

For other licenses, the following applies:

- Cloudera Express- The feature is not available.
- Cloudera Enterprise Data Hub Edition Trial - The feature is available until you end the trial or the trial license expires.

To obtain a license for Cloudera Enterprise, fill in this [form](#) or call 866-843-7207. After you install a Cloudera Enterprise license, the feature will be available.

## Backup and Disaster Recovery Overview

Cloudera Manager provides an integrated, easy-to-use management solution for enabling data protection in the Hadoop platform. Cloudera Manager provides rich functionality aimed towards replicating data stored in HDFS and accessed through Hive across data centers for disaster recovery scenarios. When critical data is stored on HDFS, Cloudera Manager provides the necessary capabilities to ensure that the data is available at all times, even in the face of the complete shutdown of a data center.

Cloudera Manager also provides the ability to schedule, save and (if needed) restore snapshots of HDFS directories and HBase tables.

Cloudera Manager provides key capabilities that are fully integrated into the Cloudera Manager Admin Console:

- **Select** - Choose the key datasets that are critical for your business operations.
- **Schedule** - Create an appropriate schedule for data replication and/or snapshots – trigger replication and snapshots as frequently as is appropriate for your business needs.
- **Monitor** - Track progress of your snapshots and replication jobs through a central console and easily identify issues or files that failed to be transferred.
- **Alert** - Issue alerts when a snapshot or replication job fails or is aborted so that the problem can be diagnosed expeditiously.

Replication capabilities work seamlessly across Hive and HDFS – replication can be setup on files or directories in the case of HDFS and on tables in the case of Hive—without any manual translation of Hive datasets into HDFS datasets or vice-versa. Hive metastore information is also replicated which means that the applications that depend upon the table definitions stored in Hive will work correctly on the replica side as well as the source side as table definitions are updated.

Built on top of a hardened version of `distcp`—the replication uses the scalability and availability of MapReduce and YARN to parallelize the copying of files using a specialized MapReduce job or YARN application that diffs and transfers only changed files from each Mapper to the replica side efficiently and quickly.

Also available is the ability to do a “Dry Run” to verify configuration and understand the cost of the overall operation before actually copying the entire dataset.

### Port Requirements

You must ensure that the following ports are open and accessible across clusters to allow communication between the source and destination Cloudera Manager servers and the HDFS, Hive, MapReduce, and YARN hosts:

- Cloudera Manager Admin Console port: Default is 7180.

- HDFS NameNode port: Default is 8020.
- HDFS DataNode port: Default is 50010.
- WebHDFS port: Default is 50070.

See [Ports](#) for more information, including how to verify the current values for these ports.

## Data Replication

Cloudera Manager provides rich functionality for replicating data (stored in HDFS or accessed through Hive) across data centers. When critical data is stored on HDFS, Cloudera Manager provides the necessary capabilities to ensure that the data is available at all times, even in the face of the complete shutdown of a data center.

For recommendations on using data replication and Sentry authorization, see [Configuring Sentry to Enable BDR Replication](#).

In Cloudera Manager 5, replication is supported between CDH 5 or CDH 4 clusters. In Cloudera Manager 5, support for HDFS and Hive replication is as follows.

### Supported Replication Scenarios

- **HDFS and Hive**
  - Cloudera Manager 4 with CDH 4 to Cloudera Manager 5 with CDH 4
  - Cloudera Manager 5 with CDH 4 to Cloudera Manager 4.7.3 or later with CDH 4
  - Cloudera Manager 5 with CDH 4 to Cloudera Manager 5 with CDH 4
  - Cloudera Manager 5 with CDH 5 to Cloudera Manager 5 with CDH 5
  - Cloudera Manager 4 or 5 with CDH 4.4 or later to Cloudera Manager 5 with CDH 5
  - Cloudera Manager 5 with CDH 5 to Cloudera Manager 5 with CDH 4.4 or later.
- **SSL**
  - Between CDH 5.0 with SSL and CDH 5.0 with SSL.
  - Between CDH 5.0 with SSL and CDH 5.0 without SSL.
  - From a CDH 5.1 source cluster with SSL and YARN.

### Unsupported Replication Scenarios

- **HDFS and Hive**
  - Cloudera Manager 5 with CDH 5 as the *source*, and Cloudera Manager 4 with CDH 4 as the *target*.
  - Between Cloudera Enterprise and any Cloudera Manager free edition: Cloudera Express, Cloudera Standard, Cloudera Manager Free Edition.
  - Between CDH 5 and CDH 4 (in either direction) where the replicated data includes a directory that contains a large number of files or subdirectories (several hundreds of thousands of entries), causing out-of-memory errors. This is because of limitations in the WebHDFS API. The workaround is to increase the heap size as follows:
    1. On the target Cloudera Manager instance, go to the HDFS service page.
    2. Click the **Configuration** tab.
    3. Expand the **Service-Wide** category.
    4. Click **Advanced** > **HDFS Replication Advanced Configuration Snippet**.
    5. Increase the heap size by adding a key-value pair, for instance, `HADOOP_CLIENT_OPTS=-Xmx1g`. In this example, `1g` sets the heap size to 1 GB. This value should be adjusted depending on the number of files and directories being replicated.

- Replication involving HDFS data from CDH 5 HA to CDH 4 clusters or CDH 4 HA to CDH5 clusters will fail if a NameNode failover happens during replication. This is because of limitations in the CDH WebHDFS API.

### ■ HDFS

- Between a source cluster that has encryption over-the-wire enabled and a target cluster running CDH 4.0. This is because the CDH 4 client is used for replication in this case, and it does not support this.
- From CDH 5 to CDH 4 where there are URL-encoding characters such as % in file and directory names. This is because of a bug in the CDH 4 WebHDFS API.
- HDFS replication does not work from CDH 5 to CDH 4 with different realms when using older JDK versions. Use JDK 7 or upgrade to JDK6u34 or later on the CDH 4 cluster to avoid this issue.
- Replication for HDFS paths with encryption-at-rest enabled is not currently supported.

### ■ Hive

- *With* data replication, between a *source* cluster that has encryption enabled and a *target* cluster running CDH 4. This is because the CDH 4 client used for replication does not support encryption.
- *Without* data replication, between a *source* cluster running CDH 4 and a *target* cluster that has encryption enabled.
- Between CDH 4.2 or later and CDH 4, if the Hive schema contains views.
- With the same cluster as both source and destination
- Replication from CDH 4 to CDH 5 HA can fail if a NameNode failover happens during replication.
- Hive replication from CDH 5 to CDH 4 with different realms with older JDK versions, if data replication is enabled (since this involves HDFS replication). Use JDK 7 or upgrade to JDK6u34 or later on the CDH 4 cluster to avoid this issue.
- Hive replication from CDH 4 to CDH 5 with different realms with older JDK versions (even without data replication enabled). Use JDK 7 or upgrade to JDK6u34 or later on the CDH 4 cluster to avoid this issue.
- Replication for Hive data from HDFS paths with encryption-at-rest enabled is not currently supported.
- Cloudera Manager 5.2 only supports replication of Impala UDFs if running CDH 5.2 or later. In clusters running CM5.2 and a CDH version earlier than 5.2 that include Impala User-Defined Functions (UDFs), Hive replication will succeed, but replication of the Impala UDFs will be skipped.

■ **Note:** If the `hadoop.proxyuser.hive.groups` configuration has been changed to restrict access to the Hive Metastore Server to certain users/groups only, then the `hdfs` group or a group containing the `hdfs` user must also be included in the list of groups specified in order for Hive replication to work. This can be specified either on the Hive service as an override, or in the core-site HDFS configuration. This note applies to configuration settings on both the source and target clusters.

### ■ SSL

- From a CDH 4.x source cluster with SSL.
- From CDH 5.0 source cluster with SSL and YARN (because of a YARN bug).
- Between CDH 5.0 with SSL and CDH 4.x.

## Designating a Replication Source

Required Role: **Cluster Administrator** **Full Administrator**

The Cloudera Manager Server that you are logged in to will be treated as the destination of replications setup via that Cloudera Manager. From the Admin Console of this destination Cloudera Manager, you can designate a peer Cloudera Manager Server which will be treated as a source of HDFS and Hive data for replication.

## Configuring a Peer Relationship

1. Navigate to the **Peers** page by selecting **Administration > Peers**. The Peers page displays. If there are no existing peers, you will see only an **Add Peer** button in addition to a short message. If you have existing peers, they are listed in the Peers list.
2. Click the **Add Peer** button.
3. In the Add Peer pop-up, provide a name, the URL (including the port) of the Cloudera Manager Server that will act as the source for the data to be replicated, and the login credentials for that server.

- **Important:** The login credentials on the source server must be those of either a *User Administrator* or a *Full Administrator*.

Cloudera recommends that SSL be used and a warning is shown if the URL scheme is `http` instead of `https`.

4. Click the **Add Peer** button in the pop-up to create the peer relationship. The peer is added to the Peers list.
5. To test the connectivity between your Cloudera Manager Server and the peer, select **Actions > Test Connectivity**.

## Modifying Peers

1. Navigate to the **Peers** page by selecting **Administration > Peers**. The Peers page displays. If there are no existing peers, you will see only an **Add Peer** button in addition to a short message. If you have existing peers, they are listed in the Peers list.
2. Choose an action and follow the procedure:
  - **Edit**
    1. From the **Actions** menu for the peer, select **Edit**.
    2. Make your changes.
    3. Click **Update Peer** to save your changes.
  - **Delete** - From the **Actions** menu for the peer, select **Delete**.

## HBase Replication

If your data is already in an HBase cluster, replication is useful for getting the data into additional HBase clusters. In HBase, cluster replication refers to keeping one cluster state synchronized with that of another cluster, using the write-ahead log (WAL) of the source cluster to propagate the changes. Replication is enabled at column family granularity. Before enabling replication for a column family, create the table and all column families to be replicated, on the destination cluster.

Cluster replication uses a master-push methodology. An HBase cluster can be a source (also called *master* or *active*, meaning that it is the originator of new data), a destination (also called *slave* or *passive*, meaning that it receives data via replication), or can fulfill both roles at once. Replication is asynchronous, and the goal of replication is consistency.

When data is replicated from one cluster to another, the original source of the data is tracked with a cluster ID, which is part of the metadata. In CDH 5, all clusters that have already consumed the data are also tracked. This prevents replication loops.

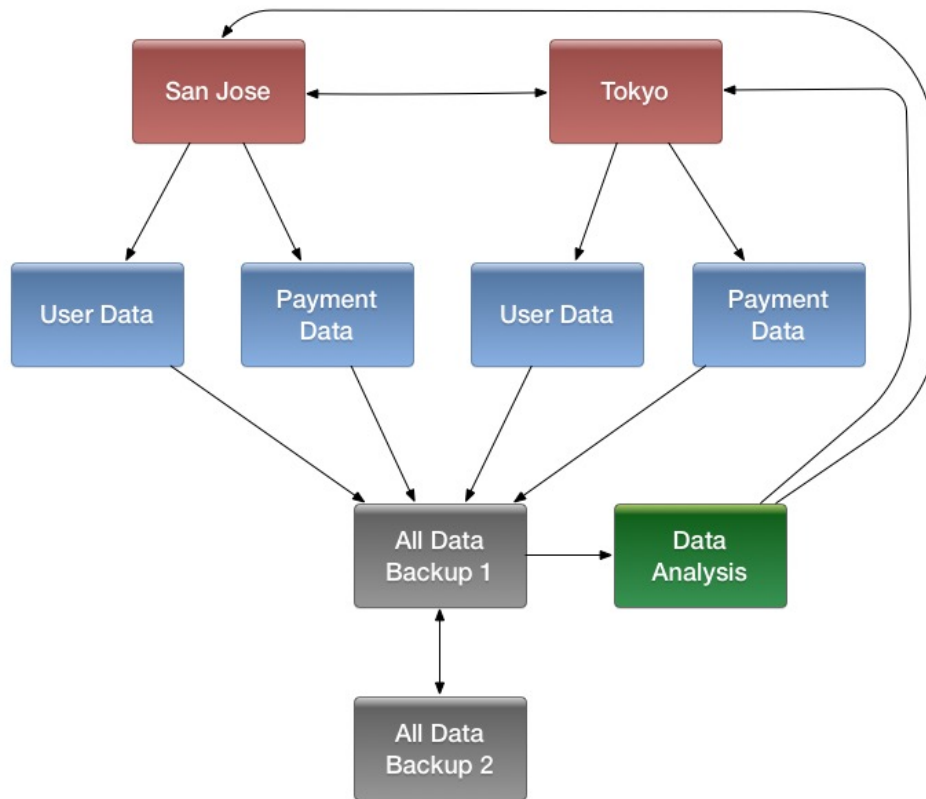
## Common Replication Topologies

- **Note:** Previously, terms such as *master-master*, *master-slave*, and *cyclic* were used to describe replication relationships in HBase. These terms were confusing and have been replaced by discussions about cluster topologies appropriate for different scenarios.

- A central source cluster might propagate changes to multiple destination clusters, for failover or due to geographic distribution.
- A source cluster might push changes to a destination cluster, which might also push its own changes back to the original cluster.

## Backup and Disaster Recovery

- Many different low-latency clusters might push changes to one centralized cluster for backup or resource-intensive data-analytics jobs. The processed data might then be replicated back to the low-latency clusters.
- Multiple levels of replication can be chained together to suit your needs. The following diagram shows a hypothetical scenario. Use the arrows to follow the data paths.



At the top of the diagram, the `San Jose` and `Tokyo` clusters, shown in red, replicate changes to each other, and each also replicates changes to a `User Data` and a `Payment Data` cluster.

Each cluster in the second row, shown in blue, replicates its changes to the `All Data Backup 1` cluster, shown in grey. The `All Data Backup 1` cluster replicates changes to the `All Data Backup 2` cluster (also shown in grey), as well as the `Data Analysis` cluster (shown in green). `All Data Backup 2` also propagates any of its own changes back to `All Data Backup 1`.

The `Data Analysis` cluster runs MapReduce jobs on its data, and then pushes the processed data back to the `San Jose` and `Tokyo` clusters.

### Points to Note about Replication

- The timestamps of the replicated HLog entries are kept intact. In case of a collision (two entries identical as to row key, column family, column qualifier, and timestamp) only the entry arriving later will be read.
- Increment Column Values (ICVs) are treated as simple puts when they are replicated. In the master-master case, this may be undesirable, creating identical counters that overwrite one another. (See <https://issues.apache.org/jira/browse/HBase-2804>.)
- Make sure the master and slave clusters are time-synchronized with each other. Cloudera recommends you use Network Time Protocol (NTP).
- Some changes are not replicated and must be propagated through other means, such as [Snapshots](#) or [CopyTable](#).
  - Data that existed in the master before replication was enabled.

- Operations that bypass the WAL, such as when using BulkLoad or API calls such as `writeToWal(false)`.
- Table schema modifications.

## Requirements

Before configuring replication, make sure your environment meets the following requirements:

- You must manage ZooKeeper yourself. It must not be managed by HBase, and must be available throughout the deployment.
- Each host in both clusters must be able to reach every other host, including those in the ZooKeeper cluster.
- Both clusters must be running the same major version of CDH; for example CDH 4 or CDH 5.
- Every table that contains families that are scoped for replication must exist on each cluster and have exactly the same name.
- HBase version 0.92 or greater is required for multiple slaves, master-master, or cyclic replication..

## Deploying HBase Replication

Follow these steps to enable replication from one cluster to another.

1. Configure and start the source and destination clusters. Create tables with the same names and column families on both the source and destination clusters, so that the destination cluster knows where to store data it receives. All hosts in the source and destination clusters should be reachable to each other.
2. On the source cluster, enable replication in Cloudera Manager, or by setting `hbase.replication` to `true` in `hbase-site.xml`.
3. On the source cluster, in HBase Shell, add the destination cluster as a peer, using the `add_peer` command. The syntax is as follows:

```
add_peer 'ID' 'CLUSTER_KEY'
```

The ID must be a short integer. To compose the `CLUSTER_KEY`, use the following template:

```
hbase.zookeeper.quorum:hbase.zookeeper.property.clientPort:zookeeper.znode.parent
```

If both clusters use the same ZooKeeper cluster, you must use a different **zookeeper.znode.parent**, because they cannot write in the same folder.

4. On the source cluster, configure each column family to be replicated by setting its `REPLICATION_SCOPE` to 1, using commands such as the following in HBase Shell.

```
hbase> disable 'example_table'
hbase> alter 'example_table', {NAME => 'example_family', REPLICATION_SCOPE => '1'}
hbase> enable 'example_table'
```

5. Verify that replication is occurring by examining the logs on the source cluster for messages such as the following.

```
Considering 1 rs, with ratio 0.1
Getting 1 rs from peer cluster # 0
Choosing peer 10.10.1.49:62020
```

6. To verify the validity of replicated data, use the included `VerifyReplication` MapReduce job on the source cluster, providing it with the ID of the replication peer and table name to verify. Other options are available, such as a time range or specific families to verify.

The command has the following form:

```
hbase org.apache.hadoop.hbase.mapreduce.replication.VerifyReplication
[--starttime=timestamp] [--stoptime=timestamp] [--families=comma separated list
of families] <peerId> <tablename>
```

## Backup and Disaster Recovery

The `VerifyReplication` command prints `GOODROWS` and `BADROWS` counters to indicate rows that did and did not replicate correctly.

### Guidelines for Replication across Three or More Clusters

When configuring replication among three or more clusters, Cloudera recommends you enable `KEEP_DELETED_CELLS` on column families in the slave cluster, where `REPLICATION_SCOPE=1` in the master cluster. The following commands show how to enable this configuration using HBase Shell.

- On the master:

```
create 't1',{NAME=>'f1', REPLICATION_SCOPE=>1}
```

- On the slave:

```
create 't1',{NAME=>'f1', KEEP_DELETED_CELLS=>'true'}
```

### Disabling Replication at the Peer Level

Use the command `disable_peer (<peerID>)` to disable replication for a specific peer. This will stop replication to the peer, but the logs will be kept for future reference.

To re-enable the peer, use the command `enable_peer(<"peerID">)`. Replication resumes where it was stopped.

#### Examples:

- To disable peer 1:

```
disable_peer("1")
```

- To re-enable peer 1:

```
enable_peer("1")
```

If you disable replication, and then later decide to enable it again, you must manually remove the old replication data from ZooKeeper by deleting the contents of the replication queue within the `/hbase/replication/rs/znode`. If you fail to do so, and you re-enable replication, the master cannot reassign previously-replicated regions. Instead, you will see logged errors such as the following:

```
2015-04-20 11:05:25,428 INFO
org.apache.hadoop.hbase.replication.ReplicationQueuesZKImpl:
  Won't transfer the queue, another RS took care of it because of: KeeperErrorCode
  = NoNode for
  /hbase/replication/rs/c856fqz.example.com,60020,1426225601879/lock
```

### Stopping Replication in an Emergency

If replication is causing serious problems, you can stop it while the clusters are running.

- **Warning:** Do this only in case of a serious problem; it may cause data loss.

#### To stop replication in an emergency:

Open the shell on the master cluster and use the `stop_replication` command. For example:

```
hbase(main):001:0> stop_replication
```

Already queued edits will be replicated after you use the `stop_replication` command, but new entries will not.



To start replication again, use the `start_replication` command.

## Initiating Replication When Data Already Exists

You may need to start replication from some point in the past. For example, suppose you have a primary HBase cluster in one location and are setting up a disaster-recovery (DR) cluster in another. To initialize the DR cluster, you need to copy over the existing data from the primary to the DR cluster, so that when you need to switch to the DR cluster you have a full copy of the data generated by the primary cluster. Once that is done, replication of new data can proceed as normal.

To start replication from an earlier point in time, run a `copyTable` command (defining the start and end timestamps), while enabling replication. Proceed as follows:

1. Start replication and note the timestamp.
2. Run the `copyTable` command with an end timestamp equal to the timestamp you noted in the previous step.

- **Note:** Because replication starts from the current WAL, some key values may be copied to the slave cluster by both the replication and the `copyTable` job. This is not a problem because this is an idempotent operation (one that can be applied multiple times without changing the result).

## Replicating Pre-existing Data in a Master-Master Deployment

In the case of master-master replication, run the `copyTable` job before starting the replication. (If you start the job after enabling replication, the second master will re-send the data to the first master, because `copyTable` does not edit the `clusterId` in the mutation objects. Proceed as follows:

1. Run the `copyTable` job and note the start timestamp of the job.
2. Start replication.
3. Run the `copyTable` job again with a start time equal to the start time you noted in step 1.

This results in some data being pushed back and forth between the two clusters; but it minimizes the amount of data.

## Configuring Secure HBase Replication

If you want to make HBase Replication secure, follow the instructions under [HBase Authentication](#).

## Caveats

- Two variables govern replication: `hbase.replication` as described above under [Deploying HBase Replication](#) on page 271, and a replication znode. Stopping replication (using `stop_replication` as above) sets the znode to `false`. Two problems can result:
  - If you add a new `RegionServer` to the master cluster while replication is stopped, its current log will not be added to the replication queue, because the replication znode is still set to `false`. If you restart replication at this point (using `start_replication`), entries in the log will not be replicated.
  - Similarly, if a logs rolls on an existing `RegionServer` on the master cluster while replication is stopped, the new log will not be replicated, because the replication znode was set to `false` when the new log was created.
- In the case of a long-running, write-intensive workload, the slave cluster may become unresponsive if its meta-handlers are blocked while performing the replication. CDH 5 provides three properties to deal with this problem:
  - `hbase.regionserver.replication.handler.count` - the number of replication handlers in the slave cluster (default is 3). Replication is now handled by separate handlers in the slave cluster to avoid the above-mentioned sluggishness. Increase it to a high value if the ratio of master to slave `RegionServers` is high.

- `replication.sink.client.retries.number` - the number of times the HBase replication client at the sink cluster should retry writing the WAL entries (default is 1).
- `replication.sink.client.ops.timeout` - the timeout for the HBase replication client at the sink cluster (default is 20 seconds).
- For namespaces, tables, column families, or cells with associated ACLs, the ACLs themselves are not replicated. The ACLs need to be re-created manually on the target table. This behavior opens up the possibility for the ACLs could be different in the source and destination cluster.

## HDFS Replication

Required Role: **BDR Administrator** **Full Administrator**

HDFS replication enables you to copy (replicate) your HDFS data from one HDFS service to another, keeping the data set on the *target* service synchronized with the data set on the *source* service, based on a user-specified replication schedule. The target service needs to be managed by the Cloudera Manager Server where the replication is being set up, and the source service could either be managed by that same server or by a peer Cloudera Manager Server.

### Configuring Replication of HDFS Data

1. Verify that your cluster conforms to the [supported replication scenarios](#).
2. If the source cluster is managed by a different Cloudera Manager server from the target cluster, [configure a peer relationship](#).
3. Do one of the following:
  - From the **Backup** tab, select **Replications**.
  - From the **Clusters** tab, go to the HDFS service and select the **Replication** tab.

The Schedules tab of the Replications page displays.

4. Click the **Schedule HDFS Replication** link.
5. Select the source HDFS service from the HDFS services managed by the peer Cloudera Manager Server or the HDFS services managed by the Cloudera Manager Server whose Admin Console you are logged into.
6. Enter the path to the directory (or file) you want to replicate (the source).
7. Select the target HDFS service from the HDFS services managed by the Cloudera Manager Server whose Admin Console you are logged into.
8. Enter the path where the target files should be placed.
9. Select a schedule. You can have it run immediately, run once at a scheduled time in the future, or at regularly scheduled intervals. If you select **Once** or **Recurring** you are presented with fields that let you set the date and time and (if appropriate) the interval between runs.
10. If you want to modify the parameters of the job, click **More Options**. Here you will be able to change the following parameters:
  - **MapReduce Service** - The MapReduce or YARN service to use.
  - **Scheduler Pool** - The scheduler pool to use.
  - **Run as** - The user that should run the job. By default this is `hdfs`. If you want to run the job as a different user, you can enter that here. If you are using Kerberos, you *must* provide a user name here, and it must be one with an ID greater than 1000. Verify that the user running the job has a home directory, `/user/<username>`, owned by `username:supergroup` in HDFS.
  - **Log path** - An alternative path for the logs.
  - **Maximum map slots** and **Maximum bandwidth** - Limits for the number of map slots and for bandwidth per mapper. The defaults are unlimited.
  - **Abort on error** - Whether to abort the job on an error (default is not to do so). This means that files copied up to that point will remain on the destination, but no additional files will be copied.
  - **Replication Strategy** - Whether file replication tasks should be distributed among the mappers statically or dynamically (the default is static). The static replication strategy distributes file replication tasks among

the mappers up front statically, trying to achieve a uniform distribution based on the file sizes. The dynamic replication strategy distributes file replication tasks in small sets to the mappers, and as each mapper is done processing its set of tasks, it dynamically picks up and processes the next unallocated set of tasks.

- **Skip Checksum Checks** - Whether to skip checksum checks (the default is to perform them). If checked, checksum validation will not be performed.
- **Delete policy** - Whether files that were removed on the source should also be deleted from the target directory. This policy also determines the handling of files that exist in the target location but are unrelated to the source. There are three options:
  - **Keep deleted files** - Retains the destination files even when they no longer exist at the source (this is the default).
  - **Delete to trash** - If the HDFS trash is enabled, files will be moved to the trash folder.
  - **Delete permanently** - Uses least amount of space, but should be used with caution.
- **Preserve** - Whether to preserve the block size, replication count, and permissions as they exist on the source file system, or to use the settings as configured on the target file system. The default is to preserve these settings as on the source.

▪ **Note:** To preserve permissions, you must be running as a superuser. You can use the "Run as" option to ensure that is the case.

- **Alerts** - Whether to generate alerts for various state changes in the replication workflow. You can alert on failure, on start, on success, or when the replication workflow is aborted.

#### 11. Click **Save Schedule**.

To specify additional replication tasks, select **Create > HDFS Replication**.


A replication task appears in the All Replications list, with relevant information about the source and target locations, the timestamp of the last job, and the next scheduled job (if there is a recurring schedule). A scheduled job will show a calendar icon to the left of the task specification. If the task is scheduled to run once, the calendar icon will disappear after the job has run.

Only one job corresponding to a replication schedule can occur at a time; if another job associated with that same replication schedule starts before the previous one has finished the second one is canceled.

From the **Actions** menu for a replication task, you can:

- Test the replication task without actually transferring data ("Dry Run" )
- Edit the task configuration
- Run the task (immediately)
- Delete the task
- Disable or enable the task (if the task is on a recurring schedule). When a task is disabled, instead of the calendar icon you will see a Stopped icon, and the job entry will appear in gray.

### Viewing Replication Job Status

- While a job is in progress, the calendar icon turns into spinner, and each stage of the replication task is indicated in the message after the replication specification.
- If the job is successful, the number of files copied is indicated. If there have been no changes to a file at the source since the previous job, then that file will *not* be copied. As a result, after the initial job, only a subset of the files may actually be copied, and this will be indicated in the success message.
- If the job fails, a  icon displays.
- To view more information about a completed job, click the task row in the Replications list. This displays sub-entries for each past job.
- To view detailed information about a past job, click the entry for that job. This opens another sub-entry that shows:
  - A result message

## Backup and Disaster Recovery

- The start and end time of the job.
  - A link to the command details for that replication job.
  - Details about the data that was replicated.
- When viewing a sub-entry, you can dismiss the sub-entry by clicking anywhere in its parent entry, or by clicking the return arrow icon at the top left of the sub-entry area.

## Hive Replication

Required Role: **BDR Administrator** **Full Administrator**

Hive replication enables you to copy (replicate) your Hive metastore and data from one cluster to another and keep the Hive metastore and data set on the target cluster synchronized with the source based on a user specified replication schedule. The target cluster needs to be managed by the Cloudera Manager Server where the replication is being set up and the source cluster could either be managed by that same server or by a peer Cloudera Manager Server.

- **Note:** If the `hadoop.proxyuser.hive.groups` configuration has been changed to restrict access to the Hive Metastore Server to certain users/groups only, then the `hdfs` group or a group containing the `hdfs` user must also be included in the list of groups specified in order for Hive replication to work. This can be specified either on the Hive service as an override, or in the core-site HDFS configuration. This note applies to configuration settings on both the source and target clusters.

## Configuring Replication of Hive Data

- **Note:** In CDH 5.2.0 Hive introduces permanent UDFs. JARs for these UDFs are stored in HDFS at a user defined location. If you are replicating Hive data you should also replicate this directory.

1. Verify that your cluster conforms to the [supported replication scenarios](#).
2. If the source cluster is managed by a different Cloudera Manager server from the target cluster, [configure a peer relationship](#).
3. Do one of the following:
  - From the **Backup** tab, select **Replications**.
  - From the **Clusters** tab, go to the Hive service and select the **Replication** tab.

The Schedules tab of the Replications page displays.

4. Click the **Schedule Hive Replication** link.
5. Select the Hive service from one managed by the local Cloudera Manager Server or from one of the Hive services managed by the peer Cloudera Manager Server to be the source of the replicated data.
6. Leave **Replicate All** checked to replicate all the Hive metastore databases from the source. To replicate only selected databases, uncheck this option and enter the database name(s) and tables you want to replicate.
  - You can specify multiple data bases and tables using the plus symbol to add more rows to the specification.
  - You can specify multiple databases on a single line by separating their names with the "|" character. For example: `mydbname1 | mydbname2 | mydbname3`.
  - Regular expressions can be used in either database or table fields. For example:

Regular Expression	Result
<code>[\w] . +</code>	Any database/table name
<code>(?!myname\b) . +</code>	Any database/table except the one named "myname"
<code>db1   db2 [\w_]+</code>	Get all tables of the db1 and db2 databases

Regular Expression	Result
db1 [\w_]+	Alternate way to get all tables of the db1 and db2 databases
Click the "+" button and then enter	
db2 [\w_]+	

7. Select the target destination. If there is only one Hive service managed by Cloudera Manager available as a target, then this will be specified as the target. If there are more than one Hive services managed by this Cloudera Manager, select from among them.
8. Select a schedule. You can have it run immediately, run once at a scheduled time in the future, or at regularly scheduled intervals. If you select **Once** or **Recurring** you are presented with fields that let you set the date and time and (if appropriate) the interval between runs.
9. Uncheck the **Replicate HDFS Files** checkbox to skip replicating the associated data files.
10. Uncheck the **Replicate Impala Metadata** checkbox to skip replicating Impala metadata. (This option is checked by default.)
11. Use the **More Options** section to specify an export location, modify the parameters of the MapReduce job that will perform the replication, and other options. Here you will be able to select a MapReduce service (if there is more than one in your cluster) and change the following parameters:
  - By default, Hive metadata is exported to a default HDFS location (`/user/${user.name}/.cm/hive`) and then imported from this HDFS file to the target Hive metastore. The default HDFS location for this export file can be overridden by specifying a path in the **Export Path** field.
  - The **Force Overwrite** option, if checked, forces overwriting data in the target metastore if there are incompatible changes detected. For example, if the target metastore was modified and a new partition was added to a table, this option would force deletion of that partition, overwriting the table with the version found on the source.

▪ **Important:** If the **Force Overwrite** option is not set and the Hive replication process detects incompatible changes on the source cluster, Hive replication will fail. This situation may arise especially with recurring replications, where the metadata associated with an existing database or table on the source cluster changes over time.

- By default, Hive's HDFS data files (say, `/user/hive/warehouse/db1/t1`) are replicated to a location relative to "/" (in this example, to `/user/hive/warehouse/db1/t1`). To override the default, enter a path in the Destination field. For example, if you enter a path such as `/ReplicatedData`, then the data files would be replicated to `/ReplicatedData/user/hive/warehouse/db1/t1`.
- Select the MapReduce service to use for this replication (if there is more than one in your cluster). The user is set in the **Run As** option.
- To specify the user that should run the MapReduce job, use the **Run As** option. By default MapReduce jobs run as `hdfs`. If you want to run the MapReduce job as a different user, you can enter that here. If you are using Kerberos, you *must* provide a user name here, and it must be one with an ID greater than 1000.
- An alternative path for the logs.
- Limits for the number of map slots and for bandwidth per mapper. The defaults are unlimited.
- Whether to abort the job on an error (default is not to abort the job). Check the checkbox to enable this. This means that files copied up to that point will remain on the destination, but no additional files will be copied.
- Whether the file replication strategy should be static or dynamic (default is static). The static replication strategy distributes file replication tasks among the mappers up front statically, trying to achieve a uniform distribution based on the file sizes. The dynamic replication strategy distributes file replication tasks in small sets to the mappers, and as each mapper is done processing its set of tasks, it dynamically picks up and processes the next unallocated set of tasks.

- Whether to skip checksum checks (default is to perform them).
- Whether files that were removed on the source should also be deleted from the target directory. There are three options: keep deleted files (this is the default), delete the files to the HDFS trash, or delete them permanently.
- Whether to preserve the block size, replication count, and permissions as they exist on the source file system, or to use the settings as configured on the target file system. The default is to preserve these settings as on the source.

▪ **Note:** If you leave the setting to preserve permissions, then you must be running as a superuser. You can use the "Run as" option to ensure that is the case.

- Whether to generate alerts for various state changes in the replication workflow. You can alert on failure, on start, on success, or when the replication workflow is aborted.

### 12. Click **Save Schedule**.

To specify additional replication tasks, select **Create > Hive Replication**.


A replication task appears in the All Replications list, with relevant information about the source and target locations, the timestamp of the last job, and the next scheduled job (if there is a recurring schedule). A scheduled job will show a calendar icon to the left of the task specification. If the task is scheduled to run once, the calendar icon will disappear after the job has run.

Only one job corresponding to a replication schedule can occur at a time; if another job associated with that same replication schedule starts before the previous one has finished the second one is canceled.

From the **Actions** menu for a replication task, you can:

- Test the replication task without actually transferring data ("Dry Run" )
- Edit the task configuration
- Run the task (immediately)
- Delete the task
- Disable or enable the task (if the task is on a recurring schedule). When a task is disabled, instead of the calendar icon you will see a Stopped icon, and the job entry will appear in gray.

## Viewing Replication Job Status

- While a job is in progress, the calendar icon turns into spinner, and each stage of the replication task is indicated in the message after the replication specification.
- If the job is successful, the number of files and tables replicated is indicated. If there have been no changes to a file at the source since the previous job, then that file will *not* be copied. As a result, after the initial job, only a subset of the files may actually be copied, and this will be indicated in the success message.
- If the job fails, a  icon displays.
- To view more information about a completed job, click the task row in the Replications list. This displays sub-entries for each past job.
- To view detailed information about a past job, click the entry for that job. This opens another sub-entry that shows:
  - A result message
  - The start and end time of the job.
  - A link to the command details for that replication job.
  - Details about the data that was replicated.
- When viewing a sub-entry, you can dismiss the sub-entry by clicking anywhere in its parent entry, or by clicking the return arrow icon at the top left of the sub-entry area.

## Impala Metadata Replication

Impala metadata replication is performed as a part of Hive replication. Impala replication is only supported between two CDH 5 clusters. The Impala and Hive services must be running on both clusters. To enable Impala metadata replication, schedule Hive replication as described in [Configuring Replication of Hive Data](#) on page 276. When performing this procedure, confirm that the **Replicate Impala Metadata** checkbox in the **Create Replication** dialog is checked.

As long as the above conditions are met, the replication of Impala metadata happens automatically as part of Hive replication. Impala metadata replication is enabled by default.

This ensures that Impala UDFs (user-defined functions) will be available on the target cluster, just as on the source cluster. As part of replicating the UDFs, the binaries in which they're defined are also replicated.

## Enabling Replication Between Clusters in Different Kerberos Realms

Required Role: **Cluster Administrator** **Full Administrator**

If you want to enable replication between clusters that reside in different Kerberos Realms, there are some additional setup steps you need to perform to ensure that the source and target clusters can communicate.

- **Note:** If either the source or target cluster is running Cloudera Manager 4.6 or later, then both clusters (source and target) must be running 4.6 or later. Cross-realm authentication does not work if one cluster is running Cloudera Manager 4.5.x and one is running Cloudera Manager 4.6 or later.

### For HDFS replication:

1. On the hosts in the *target* cluster, ensure that the `krb5.conf` file on each host has the following information:
  - The kdc information for the *source* cluster's Kerberos realm.
  - Domain/host to realm mapping for the *source* cluster NameNode hosts.
2. On the *target* cluster, through Cloudera Manager, add the realm of the *source* cluster to the Trusted Kerberos Realms configuration property.
  - a. Go to the HDFS service page and click the **Configuration** tab.
  - b. In the search field type "Trusted Kerberos" to find the Trusted Kerberos Realms property.
  - c. Enter the source cluster realm and save your changes.
3. If your Cloudera Manager is less than 5.0.1, you must restart the JobTracker to enable it to pick up the new Trusted Kerberos Realm settings. Failure to restart the JobTracker prior to the first replication attempt may cause the JobTracker to fail.

### For Hive replication:

1. Perform the steps described above on the *target* cluster, including restarting the JobTracker.
2. On the hosts in the *source* cluster, ensure that the `krb5.conf` file on each host has the following information:
  - The kdc information for the *target* cluster's Kerberos realm.
  - Domain/host to realm mapping for the *target* cluster NameNode hosts.
3. On the *source* cluster, through Cloudera Manager, add the realm of the *target* cluster to the Trusted Kerberos Realms configuration property.
  - a. Go to the HDFS service page and click the **Configuration** tab.
  - b. In the search field type "Trusted Kerberos" to find the Trusted Kerberos Realms property.
  - c. Enter the target cluster realm and save your changes.
4. It is not necessary to restart any services on the source cluster.



## Snapshots

HBase and HDFS snapshots can be created with Cloudera Manager or by using the command line.

- HBase snapshots allow you to create point-in-time backups of tables without making data copies, and with minimal impact on RegionServers. HBase snapshots are supported for clusters running CDH 4.2 or later.
- HDFS snapshots allow you to create point-in-time backups of directories or the entire filesystem without actually cloning the data. These snapshots appear on the filesystem as read-only directories that can be accessed just like any other ordinary directories. HDFS snapshots are supported for clusters running CDH 5 or later. CDH 4 does not support snapshots for HDFS.

## Cloudera Manager Snapshot Policies

Required Role: **BDR Administrator** **Full Administrator**

Cloudera Manager enables the creation of snapshot policies that define the directories or tables to be snapshotted, the intervals at which snapshots should be taken, and the number of snapshots that should be kept for each snapshot interval. For example, you can create a policy that takes both daily and weekly snapshots, and specify that 7 daily snapshots and 5 weekly snapshots should be maintained.

### Managing Snapshot Policies

- **Note:** An HDFS directory must be enabled for snapshots in order to allow snapshot policies to be created for that directory. To designate a HDFS directory as snapshottable, follow the procedure in [Enabling HDFS Snapshots](#) on page 293.

#### To create a snapshot policy:

1. Click the **Backup** tab in the top navigation bar and select **Snapshots**.

Existing snapshot policies are shown in a list organized by service. Currently running policies (if any) are shown in the **Running Policies** area.

2. To create a new policy, click **+Create**. If no policies currently exist, click the **Create snapshot policy** link. This displays the Create Snapshot Policy pop-up.
3. Select the service for which you want to create a policy from the pull-down list.
4. Provide a name for the policy and optionally a description.
5. Specify the directories or tables that should be included in the snapshot.

- For an HDFS service, select the paths of the directories that you want to include in the snapshot. The pull-down list will allow you to select only directories that have been enabled for snapshotting. If no directories have been enabled for snapshotting, a warning is displayed.

Click **+** to add another path, **-** to remove a path.

- For an HBase service, list the tables you want included in your snapshot. You can use a [Java regular expression](#) to specify a set of tables. An example is `finance.*` which will match all tables with names starting with `finance`.
6. Specify the snapshot schedule. You can schedule snapshots hourly, daily, weekly, monthly, or yearly, or any combination of those. Depending on the frequency you've selected, you can specify the time of day to take the snapshot, the day of the week, day of the month, or month of the year, and the number of snapshots to keep at each interval. Each time unit in the schedule information is shared with the time units of larger granularity. That is, the minute value is shared by all the selected schedules, hour by all the schedules for which hour is applicable, and so on. For example, if you specify that hourly snapshots are taken at the half



hour, and daily snapshots taken at the hour 20, the daily snapshot will occur at

**Schedule** ⓘ ☒ **Hourly snapshots:** Take snapshots every hour at 30 minutes. Keep 1 hourly snapshots. ✎  
☒ **Daily snapshots:** Take snapshots at 20:30. Keep 1 daily snapshots. ✎

- To select an interval, check its box. The description will then display the current schedule and the number of snapshots to retain.
- To edit the schedule (time of day, day of week and so on as relevant), and the number of snapshots to keep, click the edit icon (✎) that appears at the end of the description once you check its box. This opens an area with fields you can edit. When you have made your changes, click the **Close** button at the bottom of this area. Your changes will be reflected in the schedule description.

7. Click **More Options** to specify whether alerts should be generated for various state changes in the snapshot workflow. You can alert on failure, on start, on success, or when the snapshot workflow is aborted.

**To edit or delete a snapshot policy:**

1. Click the **Backup** tab in the top navigation bar and select **Snapshots**.
2. Click the **Actions** menu shown next to a policy and select **Edit** or **Delete**.

### Orphaned Snapshots

When a snapshot policy includes a limit on the number of snapshots to keep, Cloudera Manager checks the total number of stored snapshots each time a new snapshot is added, and automatically deletes the oldest existing snapshot if necessary. When a snapshot policy is edited or deleted, files, directories, or tables that were previously included but have now been removed from the policy may leave "orphaned" snapshots behind that will no longer be deleted automatically because they are no longer associated with a current snapshot policy. Cloudera Manager will never select these snapshots for automatic deletion because selection for deletion only occurs when the policy causes a *new* snapshot containing those files, directories, or tables to be made.

Unwanted snapshots can be deleted manually through the Cloudera Manager interface or by creating a command-line script that uses the HDFS or HBase snapshot commands. Orphaned snapshots may be hard to locate for manual deletion. Snapshot policies are automatically given a prefix `cm-auto` followed by a globally unique identifier (guid). For a specific policy, all its snapshots can be located by searching for those whose names start with the prefix `cm-auto- guid` that is unique to that policy. The prefix is prepended to the names of all snapshots created by that policy.

To avoid orphaned snapshots, delete them before editing or deleting the associated snapshot policy, or make note of the identifying name for the snapshots you want to delete. This prefix is displayed in the summary of the policy in the policy list and appears in the delete dialog. Making note of the snapshot names, including the associated policy prefix, is necessary because the prefix associated with a policy cannot be determined once the policy has been deleted, and snapshot names do not contain recognizable references to snapshot policies.

### Viewing Snapshot History

- To view the history of scheduled snapshot jobs, click a policy. This displays a list of the snapshot jobs, and their status.
- Click a snapshot job to view an expanded status for that job. (Click ⬅️ to return to the previous view.)
- From the expanded status, click the **details** link to view the details for the command. From here you can view error logs and or click **Download Result Data** to a JSON file named `summary.json` that captures information about the snapshot. For example:

```
{
  "createdSnapshotCount" : 1,
  "createdSnapshots" : [ { "creationTime" : null,
    "path" : "/user/oozie",
    "snapshotName" :
      "cm-auto-f9299438-a6eb-4f6c-90ac-5e86e5b2e283_HOURLY_2013-11-05_05-25-04",
    "snapshotPath" :
```

```
"/user/oozie/.snapshot/cm-auto-f9299438-a6eb-4f6c-90ac-5e86e5b2e283_HOURLY_2013-11-05_05-25-04"
    },
    "creationErrorCount" : 0,
    "creationErrors" : [ ],
    "deletedSnapshotCount" : 0,
    "deletedSnapshots" : [ ],
    "deletionErrorCount" : 0,
    "deletionErrors" : [ ],
    "processedPathCount" : 1,
    "processedPaths" : [ "/user/oozie" ],
    "unprocessedPathCount" : 0,
    "unprocessedPaths" : [ ]
  }
}
```

See [Managing HDFS Snapshots](#) on page 292 and [Managing HBase Snapshots](#) on page 282 for more information about managing snapshots.

### Managing HBase Snapshots

HBase snapshots can be managed using Cloudera Manager or using the command line, as described in these sections:

- [Using Cloudera Manager](#) on page 282
- [Using the Command Line](#) on page 285

#### Using Cloudera Manager

For HBase (CDH 4.2 or later or CDH 5) services, a Table Browser tab is available where you can view the HBase tables associated with a service on your cluster. From here you can view the currently saved snapshots for your tables, and delete or restore them as appropriate. From the HBase Table Browser tab you can:

- View the HBase tables that you can snapshot.
- Initiate immediate (unscheduled) snapshots of a table.
- View the list of saved snapshots currently being maintained. These may include one-off immediate snapshots, as well as scheduled policy-based snapshots.
- Delete a saved snapshot.
- Restore from a saved snapshot.
- Restore a table from a saved snapshot to a new table (Restore As).

#### Browsing HBase Tables

To browse the HBase tables to view snapshot activity:

1. From the **Clusters** tab, select your HBase service.
2. Go to the **Table Browser** tab.

#### Managing HBase Snapshots

[Required Role:](#) **BDR Administrator** **Full Administrator**

**To take a snapshot,**

1. Click a table.
2. Click **Take Snapshot**.
3. Specify the name of the snapshot, and click **Take Snapshot**.

**To delete a snapshot,** click  and select **Delete**.

**To restore a snapshot,** click  and select **Restore**.

To restore a snapshot to a new table, select **Restore As** from the menu associated with the snapshot, and provide a name for the new table.

- **Warning:** If you "Restore As" to an existing table (that is, specify a table name that already exists) the existing table will be overwritten.

## Storing HBase Snapshots on Amazon S3

With Cloudera Manager 5.2 or later and CDH 5.2 or later, HBase snapshots can be stored on the cloud storage service Amazon S3 instead of in HDFS.

- **Note:** When HBase snapshots are stored on, or restored from, Amazon S3, a MapReduce (MRv2) job is created to copy the HBase table data and metadata. For this reason, the YARN service must be running on your Cloudera Manager cluster to use this feature.

To configure HBase to store snapshots on Amazon S3, you must have the following information:

1. The *access key ID* for your Amazon S3 account.
2. The *secret access key* for your Amazon S3 account.
3. The path to the directory in Amazon S3 where you want your HBase snapshots to be stored.

### Configuring HBase in Cloudera Manager to Store Snapshots in Amazon S3

**Required Role:** **Cluster Administrator** **Full Administrator**

With the above Amazon S3 information at hand, perform the following steps in Cloudera Manager:

1. Open the HBase service page.
2. Click **Configuration**, expand **Service-Wide**, and click **Backup**.
3. Enter your Amazon S3 access key ID in the field **AWS S3 access key ID for remote snapshots**.
4. Enter your Amazon S3 secret access key in the field **AWS S3 secret access key for remote snapshots**.
5. Enter the path to the location in Amazon S3 where your HBase snapshots should be stored in the field **AWS S3 path for remote snapshots**.

- **Warning:** Do not use the Amazon S3 location defined by the path entered in **AWS S3 path for remote snapshots** for any other purpose, or directly add or delete content there. Doing so will risk corrupting the metadata associated with the HBase snapshots stored there. Use this path and Amazon S3 location only through Cloudera Manager, and only for managing HBase snapshots.

6. In a terminal window, log in to your Cloudera Manager cluster at the command line and create a `/user/hbase` directory in HDFS. Change the owner of the directory to `hbase`.

```
Example:
hdfs dfs -mkdir /user/hbase
hdfs dfs -chown hbase /user/hbase
```

### Configuring the Dynamic Resource Pool Used for Exporting and Importing Snapshots in Amazon S3

Dynamic resource pools are used to control the resources available for MapReduce jobs created for HBase snapshots on Amazon S3. By default, MapReduce jobs run against the default dynamic resource pool. To choose a different dynamic resource pool for HBase snapshots stored on Amazon S3, follow these steps:

1. Open the HBase service page.
2. Click **Configuration**, expand **Service-Wide**, and click **Backup**.
3. Enter name of a dynamic resource pool in the field **Scheduler pool for remote snapshots in AWS S3**.
4. Click **Save Changes**.

### HBase Snapshots on Amazon S3 with Kerberos Enabled

By default, when Kerberos is enabled YARN will not allow MapReduce jobs to be run by the system user `hbase`. If Kerberos is enabled on your cluster, you must perform the following steps:

1. Open the YARN service page in Cloudera Manager.
2. Click **Configuration**, expand **NodeManager Default Group**, and click **Security**.
3. In **Allowed System Users**, click the + sign and add `hbase` to the list of allowed system users.
4. Click **Save Changes**.
5. Restart the YARN service.

### Managing HBase Snapshots on Amazon S3 in Cloudera Manager

Required Role: **BDR Administrator** **Full Administrator**

To take HBase snapshots and store them on Amazon S3, perform the following steps:

1. On the HBase service page in Cloudera Manager, click the **Table Browser** tab.
2. Select a table in the Table Browser. If any recent local or remote snapshots already exist, they will be displayed on the right side.
3. In the dropdown for the selected table, click **Take Snapshot**.
4. Enter a name in the **Snapshot Name** field of the **Take Snapshot** dialog.
5. If Amazon S3 storage is configured [as described above](#), the **Take Snapshot** dialog's **Destination** section will show a choice of **Local** or **Remote S3**. Select **Remote S3**.
6. Click **Take Snapshot**.

While the **Take Snapshot** command is being executed, a local copy of the snapshot with a name beginning with `cm-tmp` followed by an auto-generated filename is displayed in the Table Browser, but this local copy is deleted as soon as the remote snapshot has been stored in Amazon S3. If the command fails without being completed, the temporary local snapshot may be left behind. This copy can be manually deleted or kept as a valid local snapshot. To store a current snapshot in Amazon S3, either execute the **Take Snapshot** command again, selecting **Remote S3** as the **Destination**, or use the HBase command-line tools to manually export the existing temporary local snapshot to Amazon S3.

### Deleting HBase Snapshots from Amazon S3

To delete a snapshot stored in Amazon S3:

1. Select the snapshot in the Table Browser.
2. Click the dropdown arrow for the snapshot.
3. Click **Delete**.

### Restoring an HBase Snapshot from Amazon S3

To restore an HBase snapshot that is stored in Amazon S3:

1. Select the table in the Table Browser.
2. Click **Restore Table**.
3. Choose **Remote S3** and select the table to restore.
4. Click **Restore**.

Cloudera Manager will create a local copy of the remote snapshot with a name beginning with `cm-tmp` followed by an auto-generated filename, and will use that local copy to restore the table in HBase. Cloudera Manager will then automatically delete the local copy. If the **Restore** command fails without being completed, the temporary copy may be left behind and will be seen in the Table Browser. In that case, delete the local temporary copy manually and re-execute the **Restore** command to restore the table from Amazon S3.

### Restoring an HBase Snapshot from Amazon S3 with a New Name

Restoring an HBase snapshot that is stored in Amazon S3 with a new name is a way of cloning the table without affecting the existing table in HBase. To do this, perform the following steps:

1. Select the table in the Table Browser.

2. Click **Restore Table From Snapshot As**.
3. In the **Restore As** dialog, enter a new name for the table in the **Restore As** field.
4. Select **Remote S3** and choose the desired snapshot in the list of available Amazon S3 snapshots.

### Managing Policies for HBase Snapshots in Amazon S3

You can configure policies to automatically create snapshots of HBase tables on an hourly, daily, weekly, monthly or yearly basis. Snapshot policies for HBase snapshots stored in Amazon S3 are configured using the same procedures as for local HBase snapshots. These procedures are described in [Cloudera Manager Snapshot Policies](#). The only additional step to perform for snapshots stored in Amazon S3 is to choose **Remote S3** in the **Destination** section of the policy management dialogs.

- **Note:** You can only configure a policy as **Local** or **Remote S3** at the time the policy is created. The setting can not be changed later. If the setting is wrong, create a new policy.

While a snapshot is being made based on a snapshot policy, as with snapshots created manually, a local copy of the snapshot is created, in this case with a name beginning `cm-auto` followed by an auto-generated filename. The temporary copy of the snapshot is displayed in the Table Browser, but this local copy is deleted as soon as the remote snapshot has been stored in Amazon S3. If the snapshot procedure fails without being completed, the temporary local snapshot may be left behind. This copy can be manually deleted or kept as a valid local snapshot. To export the HBase snapshot to Amazon S3, use the HBase command-line tools to manually export the existing temporary local snapshot to Amazon S3.

### Using the Command Line

#### About HBase Snapshots

In previous HBase releases, the only way to a backup or to clone a table was to use `CopyTable` or `ExportTable`, or to copy all the `hfiles` in HDFS after disabling the table. The disadvantages of these methods are:

- `CopyTable` and `ExportTable` can degrade region server performance.
- Disabling the table means no reads or writes; this is usually unacceptable.

HBase Snapshots allow you to clone a table without making data copies, and with minimal impact on Region Servers. Exporting the table to another cluster should not have any impact on the region servers.

### Use Cases

- Recovery from user or application errors
  - Useful because it may be some time before the database administrator notices the error

- **Note:**

The database administrator needs to schedule the intervals at which to take and delete snapshots. Use a script or your preferred management tool for this; it is not built into HBase.

- The database administrator may want to save a snapshot right before a major application upgrade or change.

- **Note:**

Snapshots are not primarily used for system upgrade protection because they would not roll back binaries, and would not necessarily be proof against bugs or errors in the system or the upgrade.

- Sub-cases for recovery:
  - Rollback to previous snapshot and merge in reverted data
  - View previous snapshots and selectively merge them into production

## Backup and Disaster Recovery

- Backup
  - Capture a copy of the database and store it outside HBase for disaster recovery
  - Capture previous versions of data for compliance, regulation, archiving
  - Export from snapshot on live system provides a more consistent view of HBase than `CopyTable` and `ExportTable`
- Audit and/or report view of data at a specific time
  - Capture monthly data for compliance
  - Use for end-of-day/month/quarter reports
- Use for Application testing
  - Test schema or application changes on like production data from snapshot and then throw away
  - For example: take a snapshot; create a new table from the snapshot content (schema plus data); manipulate the new table by changing the schema, adding and removing rows, and so on (the original table, the snapshot, and the new table remain independent of each other)
- Offload work
  - Capture, copy, and restore data to another site
  - Export data to another cluster

### Where Snapshots Are Stored

The snapshot metadata is stored in the `.hbase_snapshot` directory under the `hbase` root directory (`/hbase/.hbase-snapshot`). Each snapshot has its own directory that includes all the references to the `hfiles`, logs, and metadata needed to restore the table.

`hfiles` needed by the snapshot are in the traditional

`/hbase/data/<namespace>/<tableName>/<regionName>/<familyName>/` location if the table is still using them; otherwise they will be placed in

`/hbase/.archive/<namespace>/<tableName>/<regionName>/<familyName>/`

### Zero-copy Restore and Clone Table

From a snapshot you can create a new table (`clone` operation) or restore the original table. These two operations do not involve data copies; instead a link is created to point to the original `hfiles`.

Changes to a cloned or restored table do not affect the snapshot or (in case of a clone) the original table.

If you want to clone a table to another cluster, you need to export the snapshot to the other cluster and then execute the `clone` operation; see [Exporting a Snapshot to Another Cluster](#).

### Reverting to a Previous HBase Version

Snapshots don't affect HBase backward compatibility if they are not used.

If you do use the snapshot capability, backward compatibility is affected as follows:

- If you only take snapshots, you can still go back to a previous HBase version
- If you have used `restore` or `clone`, you cannot go back to a previous version unless the cloned or restored tables have no links (there is no automated way to check; you would need to inspect the file system manually).

### Storage Considerations

Since the `hfiles` are immutable, a snapshot consists of reference to the files that are in the table at the moment the snapshot is taken. No copies of the data are made during the snapshot operation, but copies may be made when a compaction or deletion is triggered. In this case, if a snapshot has a reference to the files to be removed, the files are moved to an archive folder, instead of being deleted. This allows the snapshot to be restored in full.

Because no copies are performed, multiple snapshots share the same `hfiles`, but in the worst case scenario, each snapshot could have different set of `hfiles` (tables with lots of updates, and compactions).

### Configuring and Enabling Snapshots

Snapshots are on by default; to disable them, set the `hbase.snapshot.enabled` property in `hbase-site.xml` to `false`:

```
<property>
  <name>hbase.snapshot.enabled</name>
  <value>
    false
  </value>
</property>
```

To enable snapshots after you have disabled them, set `hbase.snapshot.enabled` to `true`.

■ **Note:**

If you have taken snapshots and then decide to disable snapshots, you must delete the snapshots before restarting the HBase master; the HBase master will not start if snapshots are disabled and snapshots exist.

Snapshots don't affect HBase performance if they are not used.

### Shell Commands

You can manage snapshots by using the HBase shell or the HBaseAdmin Java API.

The following table shows actions you can take from the shell:

Action	Shell command	Comments
Take a snapshot of <code>tableX</code> called <code>snapshotX</code>	<pre>snapshot 'tableX', 'snapshotX'</pre>	<p>Snapshots can be taken while a table is disabled, or while a table is online and serving traffic.</p> <ul style="list-style-type: none"> <li>▪ If a table is disabled (via <code>disable &lt;table&gt;</code>) an offline snapshot is taken. This snapshot is driven by the master and fully consistent with the state when the table was disabled. This is the simplest and safest method, but it involves a service interruption since the table must be disabled to take the snapshot.</li> <li>▪ In an online snapshot, the table remains available while the snapshot is taken, and should incur minimal noticeable performance degradation of normal read/write loads. This snapshot is coordinated by the master and run on the region servers. The current implementation - simple-flush snapshots - provides no causal consistency guarantees. Despite this shortcoming, it offers the same degree of consistency as <code>CopyTable</code> and overall is a huge improvement over <code>CopyTable</code>.</li> </ul>
Restore snapshot <code>snapshotX</code> (it will replace the source table content)	<pre>restore_snapshot 'snapshotX'</pre>	<p>Restoring a snapshot attempts to replace the current version of a table with another version of the table. To run this command, you must disable the target table. The <code>restore</code> command takes a snapshot of the table (appending a timestamp code), and then essentially clones data into the original data and removes data not in the snapshot. If the operation succeeds, the target table will be enabled. Use this capability only in an emergency; see <a href="#">Restrictions</a>.</p>
List all available snapshots	<pre>list_snapshots</pre>	
List all available snapshots starting with <code>'mysnapshot_'</code> (regular expression)	<pre>list_snapshots 'my_snapshot_*'</pre>	
Remove a snapshot called <code>snapshotX</code>	<pre>delete_snapshot 'snapshotX'</pre>	
Create a new table <code>tableY</code> from a snapshot <code>snapshotX</code>	<pre>clone_snapshot 'snapshotX', 'tableY'</pre>	<p>Cloning a snapshot creates a new read/write table that can serve the data kept at the time of the snapshot. The original table and the cloned table can be modified independently without interfering – new data written to one table will not show up on the other.</p>

## Taking a Snapshot Using a Shell Script



With HBase in CDH 5.2 and newer, you can take a snapshot using an operating system shell script, such as a Bash script. This is possible because of HBase Shell's new non-interactive mode, which is described in [Accessing HBase by using the HBase Shell](#). This example Bash script illustrates how to take a snapshot in this way. This script is not production-ready, but is provided as an illustration only.

```
#!/bin/bash
# Take a snapshot of the table passed as an argument
# Usage: snapshot_script.sh table_name
# Names the snapshot in the format snapshot-YYYYMMDD

# Parse the arguments
if [ -z $1 ] || [ $1 == '-h' ]; then
    echo "Usage: $0 <table>"
    echo "        $0 -h"
    exit 1
fi

# Modify to suit your environment
export HBASE_PATH=/home/user/hbase
export DATE=`date +%Y%m%d`
echo "snapshot '$1', 'snapshot-$DATE'" | $HBASE_PATH/bin/hbase shell -n
status=$?
if [ $status -ne 0 ]; then
    echo "Snapshot may have failed: $status"
fi
exit $status
```

HBase Shell returns an exit code of 0 on success, but a non-zero exit code only indicates the possibility of failure, rather than definite failure. Therefore, your script should check to see if the snapshot was created before trying again, in the event of a reported failure.

### Exporting a Snapshot to Another Cluster

You can export any snapshot from one cluster to another. Exporting the snapshot copies the table's hfiles, logs, and the snapshot's metadata, from the source cluster to the destination cluster. You can specify the `-copy-from` option to copy from a remote cluster to the local cluster or another remote cluster. If you do not specify the `-copy-from` option, the `hbase.rootdir` in the HBase configuration is used, which means that you are exporting from the current cluster. You must specify the `-copy-to` option, to specify the destination cluster.

- **Note:** Snapshots must be enabled on the destination cluster. See [Configuring and Enabling Snapshots](#) on page 287.

The `ExportSnapshot` tool executes a MapReduce Job, similar to `distcp`, to copy files to the other cluster. It works at file-system level, so the hbase cluster can be offline.

**To copy a snapshot called `MySnapshot` to an HBase cluster `srv2 (hdfs://srv2:8082/hbase)` using 16 mappers:**

```
hbase class org.apache.hadoop.hbase.snapshot.ExportSnapshot -snapshot MySnapshot -copy-to
hdfs://srv2:8082/hbase -mappers 16
```

**To export the snapshot and change the ownership of the files during the copy:**

```
hbase class org.apache.hadoop.hbase.snapshot.ExportSnapshot -snapshot MySnapshot -copy-to
hdfs://srv2:8082/hbase -chuser MyUser -chgroup MyGroup -chmod 700 -mappers 16
```

You can also use the Java `-D` option in many tools to specify MapReduce or other configuration. properties. For example, the following command copies `MY_SNAPSHOT` to `hdfs://cluster2/hbase` using groups of 10 hfiles per mapper:

```
hbase org.apache.hadoop.hbase.snapshot.ExportSnapshot
-Dsnapshot.export.default.map.group=10 -snapshot MY_SNAPSHOT -copy-to
hdfs://cluster2/hbase
```

(The number of mappers is calculated as `TotalNumberOfHFiles/10`.)

To export from one remote cluster to another remote cluster, specify both `-copy-from` and `-copy-to` parameters. You could then reverse the direction to restore the snapshot back to the first remote cluster.

```
hbase org.apache.hadoop.hbase.snapshot.ExportSnapshot -snapshot snapshot-test -copy-from
hdfs://machine1/hbase -copy-to hdfs://machine2/my-backup
```

To specify a different name for the snapshot on the target cluster, use the `-target` option.

```
hbase org.apache.hadoop.hbase.snapshot.ExportSnapshot -snapshot snapshot-test -copy-from
hdfs://machine1/hbase -copy-to hdfs://machine2/my-backup -target new-snapshot
```

### Restrictions

- **Warning:**

**Do not use merge in combination with snapshots. Merging two regions can cause data loss if snapshots or cloned tables exist for this table.**

The merge is likely to corrupt the snapshot and any tables cloned from the snapshot. In addition, if the table has been restored from a snapshot, the merge may also corrupt the table. The snapshot may survive intact if the regions being merged are not in the snapshot, and clones may survive if they do not share files with the original table or snapshot. You can use the `SnapshotInfo` tool (see [Information and Debugging](#) on page 291) to check the status of the snapshot. If the status is `BROKEN`, the snapshot is unusable.

- All the Masters and Region Servers must be running CDH 5.
- If you have [enabled](#) the `AccessController` Coprocessor for HBase, only a global administrator can take, clone, or restore a snapshot, and these actions do not capture the ACL rights. This means that restoring a table preserves the ACL rights of the existing table, while cloning a table creates a new table that has no ACL rights until the administrator adds them.
- Do not take, clone, or restore a snapshot during a rolling restart. Snapshots rely on the Region Servers being up; otherwise the snapshot will fail.

- **Note:** This restriction also applies to rolling upgrade, which can currently be done only via Cloudera Manager.

**If you are using HBase Replication and you need to restore a snapshot:** If you are using [HBase Replication](#) the replicas will be out of synch when you restore a snapshot. Do this only in an emergency.

- **Important:**

Snapshot restore is an emergency tool; you need to disable the table and [table replication](#) to get to an earlier state, and you may lose data in the process.

If you need to restore a snapshot, proceed as follows:

1. Disable the table that is the restore target, and stop the replication
2. Remove the table from both the master and slave clusters
3. Restore the snapshot on the master cluster
4. Create the table on the slave cluster and use `CopyTable` to initialize it.

■ **Note:**

If this is not an emergency (for example, if you know that you have lost just a set of rows such as the rows starting with "xyz"), you can create a clone from the snapshot and create a MapReduce job to copy the data that you've lost.

In this case you don't need to stop replication or disable your main table.

## Snapshot Failures

Region moves, splits, and other metadata actions that happen while a snapshot is in progress will probably cause the snapshot to fail; the software detects and rejects corrupted snapshot attempts.

## Information and Debugging

You can use the `SnapshotInfo` tool to get information about a snapshot, including status, files, disk usage, and debugging information.

### Examples:

Use the `-h` option to print usage instructions for the `SnapshotInfo` utility.

```
$ hbase org.apache.hadoop.hbase.snapshot.SnapshotInfo -h
Usage: bin/hbase org.apache.hadoop.hbase.snapshot.SnapshotInfo [options]
where [options] are:
  -h|-help           Show this help and exit.
  -remote-dir        Root directory that contains the snapshots.
  -list-snapshots    List all the available snapshots and exit.
  -snapshot NAME     Snapshot to examine.
  -files             Files and logs list.
  -stats             Files and logs stats.
  -schema            Describe the snapshottable table.
```

Use the `-list-snapshots` option to list all snapshots and exit. This option is new in CDH 5.1.

```
$ hbase org.apache.hadoop.hbase.snapshot.SnapshotInfo -list-snapshots
SNAPSHOT      | CREATION TIME | TABLE NAME
snapshot-test | 2014-06-24T19:02:54 | test
```

Use the `-remote-dir` option with the `-list-snapshots` option to list snapshots located on a remote system.

```
$ hbase org.apache.hadoop.hbase.snapshot.SnapshotInfo -remote-dir
s3n://mybucket/mysnapshot-dir -list-snapshots
SNAPSHOT | CREATION TIME | TABLE NAME
snapshot-test | 2014-05-01 10:30 | myTable
```

Use the `-snapshot` option to print information about a specific snapshot.

```
$ hbase org.apache.hadoop.hbase.snapshot.SnapshotInfo -snapshot test-snapshot
Snapshot Info
-----
Name: test-snapshot
Type: DISABLED
Table: test-table
Version: 0
Created: 2012-12-30T11:21:21
*****
```

Use the `-snapshot` with the `-stats` options to display additional statistics about a snapshot.

```
$ hbase org.apache.hadoop.hbase.snapshot.SnapshotInfo -stats -snapshot snapshot-test
Snapshot Info
-----
Name: snapshot-test
Type: FLUSH
Table: test
```

```
Format: 0
Created: 2014-06-24T19:02:54

1 HFiles (0 in archive), total size 1.0k (100.00% 1.0k shared with the source table)
```

Use the `-schema` option with the `-snapshot` option to display the schema of a snapshot.

```
$ hbase org.apache.hadoop.hbase.snapshot.SnapshotInfo -schema -snapshot snapshot-test
Snapshot Info
-----
  Name: snapshot-test
  Type: FLUSH
  Table: test
  Format: 0
  Created: 2014-06-24T19:02:54

Table Descriptor
-----
'test', {NAME => 'cf', DATA_BLOCK_ENCODING => 'FAST_DIFF', BLOOMFILTER => 'ROW',
REPLICATION_SCOPE => '0',
COMPRESSION => 'GZ', VERSIONS => '1', TTL => 'FOREVER', MIN_VERSIONS => '0',
KEEP_DELETED_CELLS => 'false',
BLOCKSIZE => '65536', IN_MEMORY => 'false', BLOCKCACHE => 'true'}
```

Use the `-files` option with the `-snapshot` option to list information about files contained in a snapshot.

```
$ hbase org.apache.hadoop.hbase.snapshot.SnapshotInfo -snapshot test-snapshot -files
Snapshot Info
-----
  Name: test-snapshot
  Type: DISABLED
  Table: test-table
  Version: 0
  Created: 2012-12-30T11:21:21

Snapshot Files
-----
  52.4k test-table/02ba3a0f8964669520cf96bb4e314c60/cf/bdf29c39da2a4f2b81889eb4f7b18107
  (archive)
  52.4k test-table/02ba3a0f8964669520cf96bb4e314c60/cf/1e06029d0a2a4a709051b417aec88291
  (archive)
  86.8k test-table/02ba3a0f8964669520cf96bb4e314c60/cf/506f601e14dc4c74a058be5843b99577
  (archive)
  52.4k test-table/02ba3a0f8964669520cf96bb4e314c60/cf/5c7f6916ab724eachbcea218a713941c4
  (archive)
  293.4k test-table/02ba3a0f8964669520cf96bb4e314c60/cf/aec5e33a6564441d9bd423e31fc93abb
  (archive)
  52.4k test-table/02ba3a0f8964669520cf96bb4e314c60/cf/97782b2fbf0743edaacd8fef06ba51e4
  (archive)

6 HFiles (6 in archive), total size 589.7k (0.00% 0.0 shared with the source table)
0 Logs, total size 0.0
```

## Managing HDFS Snapshots

HDFS snapshots can be managed using Cloudera Manager or using the command line, as described in these sections:

- [Managing HDFS Snapshots Using Cloudera Manager](#) on page 292
- [Managing HDFS Snapshots Using the Command Line](#) on page 294

### Managing HDFS Snapshots Using Cloudera Manager

For HDFS (CDH 5 only) services, a File Browser tab is available where you can view the HDFS directories associated with a service on your cluster. From here you can view the currently saved snapshots for your files, and delete or restore them as appropriate. From the HDFS File Browser tab you can:

- Designate HDFS directories to be "snapshottable" so snapshots can be created for those directories.
- Initiate immediate (unscheduled) snapshots of a table.
- View the list of saved snapshots currently being maintained. These may include one-off immediate snapshots, as well as scheduled policy-based snapshots.
- Delete a saved snapshot.
- Restore an HDFS directory or file from a saved snapshot.
- Restore an HDFS directory or file from a saved snapshot to a new directory or file (Restore As)

■ **Note:** Cloudera Manager does not support snapshot operations for HDFS paths with encryption-at-rest enabled. This limitation is only for Cloudera Manager, and does not effect CDH command-line tools.

### Browsing HDFS Directories

To browse the HDFS directories to view snapshot activity:

1. From the **Clusters** tab, select your CDH 5 HDFS service.
2. Go to the **File Browser** tab.

As you browse the directory structure of your HDFS, basic information about the directory you have selected is shown at the right (owner, group, and so on).

### Enabling HDFS Snapshots

**Required Role:** **Cluster Administrator** **Full Administrator**

HDFS directories must be enabled for snapshots in order for snapshots to be created. You cannot specify a directory as part of a snapshot policy unless it has been enabled for snapshotting.

**To enable a HDFS directory for snapshots:**

1. From the **Clusters** tab, select your CDH 5 HDFS service.
2. Go to the **File Browser** tab.
3. Verify the Snapshottable Path and click **Enable Snapshots**.

■ **Note:** Once you enable snapshots for a directory, you cannot enable snapshots on any of its subdirectories. Snapshots can be taken only on directories that have snapshots enabled.

To disable snapshots for a directory that has snapshots enabled, use the **Disable Snapshots** from the drop-down menu button at the upper right. If there are existing snapshots of the directory, they must be deleted before snapshots can be disabled.

### Managing HDFS Snapshots

**Required Role:** **BDR Administrator** **Full Administrator**

If a directory has been enabled for snapshots:

- The **Take Snapshot** button is present, enabling an immediate snapshot of the directory.
- Any snapshots that have been taken are listed by the time at which they were taken, along with their names and a menu button.

**To take a snapshot,** click **Take Snapshot**, specify the name of the snapshot, and click **Take Snapshot**. The snapshot is added to the snapshot list.

**To delete a snapshot,** click  and select **Delete**.

**To restore a snapshot,** click  and select **Restore**.

For restoring HDFS data, if a MapReduce or YARN service is present in the cluster, then DistributedCopy (distcp) will be used to restore directories, increasing the speed of restoration. The restore popup for HDFS (under More Options) allows selection of either MapReduce or YARN as the MapReduce service. For files, or if a MapReduce or YARN service is not present, a normal copy will be performed. Use of distcp allows configuration of the following options for the snapshot restoration, similar to what is available when configuring a replication:

- **MapReduce Service** - The MapReduce or YARN service to use.
- **Scheduler Pool** - The scheduler pool to use.
- **Run as** - The user that should run the job. By default this is `hdfs`. If you want to run the job as a different user, you can enter that here. If you are using Kerberos, you *must* provide a user name here, and it must be one with an ID greater than 1000. Verify that the user running the job has a home directory, `/user/<username>`, owned by `username:supergroup` in HDFS.
- **Log path** - An alternative path for the logs.
- **Maximum map slots** and **Maximum bandwidth** - Limits for the number of map slots and for bandwidth per mapper. The defaults are unlimited.
- **Abort on error** - Whether to abort the job on an error (default is not to do so). This means that files copied up to that point will remain on the destination, but no additional files will be copied.
- **Skip Checksum Checks** - Whether to skip checksum checks (the default is to perform them). If checked, checksum validation will not be performed.
- **Delete policy** - Whether files that were removed on the source should also be deleted from the target directory. This policy also determines the handling of files that exist in the target location but are unrelated to the source. There are three options:
  - **Keep deleted files** - Retains the destination files even when they no longer exist at the source (this is the default).
  - **Delete to trash** - If the HDFS trash is enabled, files will be moved to the trash folder.
  - **Delete permanently** - Uses least amount of space, but should be used with caution.
- **Preserve** - Whether to preserve the block size, replication count, and permissions as they exist on the source file system, or to use the settings as configured on the target file system. The default is to preserve these settings as on the source.

▪ **Note:** To preserve permissions, you must be running as a superuser. You can use the "Run as" option to ensure that is the case.

### Managing HDFS Snapshots Using the Command Line

For information about managing snapshots using the command line, see [HDFS Snapshots](#).

# Cloudera Manager Administration

## Managing the Cloudera Manager Server and Agents

This section covers information on managing the Cloudera Manager Server and Agents that run on each host of the cluster.

### Starting, Stopping, and Restarting the Cloudera Manager Server

To start the Cloudera Manager Server:

```
$ sudo service cloudera-scm-server start
```

You can stop (for example, to perform maintenance on its host) or restart the Cloudera Manager Server without affecting the other services running on your cluster. Statistics data used by activity monitoring and service monitoring will continue to be collected during the time the server is down.

To stop the Cloudera Manager Server:

```
$ sudo service cloudera-scm-server stop
```

To restart the Cloudera Manager Server:

```
$ sudo service cloudera-scm-server restart
```

### Configuring Cloudera Manager Server Ports

**Required Role:** Full Administrator

1. Select **Administration > Settings**.
2. Under the **Ports and Addresses** category, set the following options as described below:

Setting	Description
HTTP Port for Admin Console	Specify the HTTP port to use to access the Server via the Admin Console.
HTTPS Port for Admin Console	Specify the HTTPS port to use to access the Server via the Admin Console.
Agent Port to connect to Server	Specify the port for Agents to use to connect to the Server.

3. Click **Save Changes**.
4. [Restart the Cloudera Manager Server](#).

### Moving the Cloudera Manager Server to a New Host

You can move the Cloudera Manager Server if either the Cloudera Manager database server or a current [back up](#) of the Cloudera Manager database is available. To move Cloudera Manager Server:

1. Identify a new host on which to install Cloudera Manager.
2. Install Cloudera Manager on a new host, using the method described under [Install the Cloudera Manager Server Packages](#). Do not install the other components, such as CDH and databases.
3. If the database server is not available:

- a. Install the database packages on the host that will host the restored database. This could be the same host on which you have just installed Cloudera Manager or it could be a different host. If you used the embedded PostgreSQL database, install the PostgreSQL package as described in [Embedded PostgreSQL Database](#). If you used an external MySQL, PostgreSQL, or Oracle database, reinstall following the instructions in [Cloudera Manager and Managed Service Data Stores](#).
  - b. Restore the backed up databases to the new database installation.
4. Update `/etc/cloudera-scm-server/db.properties` with the database name, database instance name, user name, and password.
  5. In `/etc/cloudera-scm-agent/config.ini` on each host, update the `server_host` property to the new hostname and restart the Agents.
  6. Start the Cloudera Manager Server. Cloudera Manager should resume functioning as it did before the failure. Because you restored the database from the backup, the server should accept the running state of the Agents, meaning it will not terminate any running processes.

The process is similar with secure clusters, though files in `/etc/cloudera-scm-server` must be restored in addition to the database. See [Cloudera Security](#).

## Starting, Stopping, and Restarting Cloudera Manager Agents

### Starting Agents

To start Agents, the `supervisord` process, and *all managed service processes*, use one of the following commands:

- **Start**

```
$ sudo service cloudera-scm-agent start
```

- **Clean Start**

```
$ sudo service cloudera-scm-agent clean_start
```

The directory `/var/run/cloudera-scm-agent` is completely cleaned out; all files and subdirectories are removed, and then the `start` command is executed. `/var/run/cloudera-scm-agent` contains on-disk running Agent state. Some Agent state is left behind in `/var/lib/cloudera-scm-agent`, but you shouldn't delete that. For further information, see [Server and Client Configuration](#) and [Process Management](#).

### Stopping and Restarting Agents

To stop or restart Agents *while leaving the managed processes running*, use one of the following commands:

- **Stop**

```
$ sudo service cloudera-scm-agent stop
```

- **Restart**

```
$ sudo service cloudera-scm-agent restart
```

### Hard Stopping and Restarting Agents

- **Warning:** The `hard_stop`, `clean_restart`, or `hard_restart` commands kill all running managed service processes on the host(s) where the command is run.

To stop or restart Agents, the `supervisord` process, and *all managed service processes*, use one of the following commands:



- **Hard Stop**

```
$ sudo service cloudera-scm-agent hard_stop
```

- **Hard Restart**

```
$ sudo service cloudera-scm-agent hard_restart
```

Hard restart is useful for the following situations:

1. You're upgrading Cloudera Manager and the `supervisord` code has changed between your current version and the new one. To properly do this upgrade you'll need to restart supervisor too.
2. `supervisord` is hung and needs to be restarted.
3. You want to clear out all running state pertaining to Cloudera Manager and managed services.

- **Clean Restart**

```
$ sudo service cloudera-scm-agent clean_restart
```

Runs `hard_stop` followed by `clean_start`.

### Checking Agent Status

To check the status of the Agent process, use the command:

```
$ sudo service cloudera-scm-agent status
```

## Configuring Cloudera Manager Agents

**Required Role:** Full Administrator

Cloudera Manager Agents can be configured globally using properties you set in the Cloudera Manager Admin Console and by setting properties in Agent configuration files.

### Configuring Agent Heartbeat and Health Status Options

You can configure the Cloudera Manager Agent heartbeat interval and timeouts to trigger changes in Agent [health](#) as follows:

1. Select **Administration > Settings**.
2. Under the **Performance** category, set the following option:

Property	Description
Send Agent Heartbeat Every	The interval in seconds between each heartbeat that is sent from Cloudera Manager Agents to the Cloudera Manager Server.  Default: 15 sec.

3. Under the **Monitoring** category, set the following options:

Property	Description
Set health status to Concerning if the Agent heartbeats fail	The number of missed consecutive heartbeats after which a <b>Concerning</b> health status is assigned to that Agent.  Default: 5.

Property	Description
Set health status to Bad if the Agent heartbeats fail	The number of missed consecutive heartbeats after which a <b>Bad</b> health status is assigned to that Agent.  Default: 10.

4. Click **Save Changes**.

Configuring the Host Parcel Directory

To configure the location of distributed parcels:

- 1. Click **Hosts** in the top navigation bar.
- 2. Click the **Configuration** tab.
- 3. Configure the value of the **Parcel Directory** property. The setting of the `parcel_dir` property in the [Cloudera Manager Agent configuration file](#) overrides this setting.
- 4. Click **Save Changes** to commit the changes.
- 5. On each host, restart the Cloudera Manager Agent:

```
$ sudo service cloudera-scm-agent restart
```

Agent Configuration File

The Cloudera Manager Agent supports different types of configuration options in the `/etc/cloudera-scm-agent/config.ini` file. You must update the configuration on each host. After changing a property, restart the Agent:

```
$ sudo service cloudera-scm-agent restart
```

Section	Property	Description
[General]	<code>server_host</code> , <code>server_port</code> , <code>listening_port</code> , <code>listening_hostname</code> , <code>listening_ip</code>	<p>Hostname and ports of the Cloudera Manager Server and Agent and IP address of the Agent. Also see <a href="#">Configuring Cloudera Manager Server Ports</a> on page 295 and <a href="#">Ports Used by Cloudera Manager and Cloudera Navigator</a>.</p> <p>The Cloudera Manager Agent configures its hostname automatically. However, if your cluster hosts are multi-homed (that is, they have more than one hostname), and you want to specify which hostname the Cloudera Manager Agent uses, you can update the <code>listening_hostname</code> property. If you want to specify which IP address the Cloudera Manager Agent uses, you can update the <code>listening_ip</code> property in the same file.</p> <p>To have a <a href="#">CNAME</a> used throughout instead of the regular hostname, an Agent can be configured to use <code>listening_hostname=CNAME</code>. In this case, the CNAME should resolve to the same IP address as the IP address of the hostname on that machine. Users doing this will find that the host inspector will report problems, but the CNAME will be used in all configurations where that's appropriate. This practice is particularly useful for users who would like clients to use <code>namenode.mycluster.company.com</code> instead of <code>machine1234.mycluster.company.com</code>. In this case, <code>namenode.mycluster</code> would be a CNAME for <code>machine1234.mycluster</code>, and the generated client</p>

Section	Property	Description
		configurations (and internal configurations as well) would use the CNAME.
	lib_dir	Directory to store Cloudera Manager Agent state that persists across instances of the agent process and system reboots. The Agent UUID is stored here.  Default: <code>/var/lib/cloudera-scm-agent</code> .
	local_filesystem_whitelist	The list of local filesystems that should always be monitored.  Default: <code>ext2,ext3,ext4</code> .
	log_file	The path to the Agent log file. If the Agent is being started via the <code>init.d</code> script, <code>/var/log/cloudera-scm-agent/cloudera-scm-agent.out</code> will also have a small amount of output (from before logging is initialized).  Default: <code>/var/log/cloudera-scm-agent/cloudera-scm-agent.log</code> .
	max_collection_wait_seconds	Maximum time to wait for all metric collectors to finish collecting data.  Default: 10 sec.
	metrics_url_timeout_seconds	Maximum time to wait when connecting to a local role's web server to fetch metrics.  Default: 30 sec.
	parcel_dir	Directory to store unpacked parcels.  Default: <code>/opt/cloudera/parcels</code> .
	supervisord_port	The supervisord port. A change takes effect the next time supervisord is restarted (not when the Agent is restarted).  Default: 19001.
	task_metrics_timeout_seconds	Maximum time to wait when connecting to a local TaskTracker to fetch task attempt data.  Default: 5 sec.
[Security]	use_tls,verify_cert_file,client_key_file,client_keypw_file,client_cert_file	Security-related configuration. See <ul style="list-style-type: none"> <li>▪ <a href="#">Level 3: Configuring TLS Authentication of Agents to the Cloudera Manager Server</a></li> <li>▪ <a href="#">Level 2: Configuring TLS Verification of Cloudera Manager Server by the Agents</a></li> <li>▪ <a href="#">Specifying the Cloudera Manager Server Certificate</a></li> <li>▪ <a href="#">Adding a Host to the Cluster</a> on page 101</li> </ul>
[Cloudera]	mgmt_home	Directory to store Cloudera Management Service files.  Default: <code>/usr/share/cmf</code> .

Section	Property	Description
[ JDBC ]	cloudera_mysql_connector_jar, cloudera_oracle_connector_jar, cloudera_postgresql_jdbc_jar	Location of JDBC drivers. See <a href="#">Cloudera Manager and Managed Service Data Stores</a> .  Default: <ul style="list-style-type: none"> <li>MySQL - /usr/share/java/mysql-connector-java.jar</li> <li>Oracle - /usr/share/java/oracle-connector-java.jar</li> <li>PostgreSQL - /usr/share/cmf/lib/postgresql-version-build.jdbc4.jar</li> </ul>

## Managing Cloudera Manager Server and Agent Logs

### Viewing Logs

To help you troubleshoot problems, you can view the Cloudera Manager Server and Agent logs. You can view these logs in the Logs page or in specific pages for the logs.

#### Viewing Cloudera Manager Server and Agent Logs in the Logs Page

1. Select **Diagnostics > Logs** on the top navigation bar.
2. Click **Select Sources** to display the log source list.
3. Uncheck the **All Sources** checkbox.
4. Check the **Cloudera Manager** checkbox to view both Agent and Server logs, or click ► to the left of Cloudera Manager, and check either the **Agent** or **Server** checkbox.
5. Click **Search**.

For more information about the Logs page, see [Logs](#).

#### Viewing the Cloudera Manager Server Log

1. Select **Diagnostics > Server Log** on the top navigation bar.

- **Note:** You can also view the Cloudera Manager Server log at `/var/log/cloudera-scm-server/cloudera-scm-server.log` on the Server host.

#### Viewing the Cloudera Manager Agent Log

1. Click the **Hosts** tab.
2. Click the link for the host where you want to see the Agent log.
3. In the **Details** panel, click the **Details** link in the **Host Agent** field.
4. Click the **Agent Log** link.

- **Note:** You can also view the Cloudera Manager Agent log at `/var/log/cloudera-scm-agent/cloudera-scm-agent.log` on the Agent hosts.

## Configuring Log Locations

### Setting the Cloudera Manager Server Log Location

By default the Cloudera Manager Server log is stored in `/var/log/cloudera-scm-server/`. If there is not enough space in that directory, you can change the location of the parent of the log directory:

1. Stop the Cloudera Manager Server:

```
$ sudo service cloudera-scm-server stop
```

2. Set the `CMF_VAR` environment variable in `/etc/default/cloudera-scm-server` to the new parent directory:

```
export CMF_VAR=/opt
```

3. Create `log/cloudera-scm_server` and `run` directories in the new parent directory and set the owner and group of all directories to `cloudera-scm`. For example, if the new parent directory is `/opt/`, do the following:

```
$ sudo su
$ cd /opt
$ mkdir log
$ chown cloudera-scm:cloudera-scm log
$ mkdir /opt/log/cloudera-scm-server
$ chown cloudera-scm:cloudera-scm log/cloudera-scm-server
$ mkdir run
$ chown cloudera-scm:cloudera-scm run
```

4. Restart the Cloudera Manager Server:

```
$ sudo service cloudera-scm-server start
```

### Setting the Cloudera Manager Agent Log Location

By default the Cloudera Manager Agent log is stored in `/var/log/cloudera-scm-agent/`. If there is not enough space in that directory, you can change the location of the log file:

1. Set the `log_file` property in the Cloudera Manager Agent [configuration file](#):

```
log_file=/opt/log/cloudera-scm-agent/cloudera-scm-agent.log
```

2. Create `log/cloudera-scm_agent` directories and set the owner and group to `cloudera-scm`. For example, if the log is stored in `/opt/log/cloudera-scm-agent`, do the following:

```
$ sudo su
$ cd /opt
$ mkdir log
$ chown cloudera-scm:cloudera-scm log
$ mkdir /opt/log/cloudera-scm-agent
$ chown cloudera-scm:cloudera-scm log/cloudera-scm-agent
```

3. Restart the Agent:

```
$ sudo service cloudera-scm-agent restart
```

## Changing Hostnames

**Required Role:** Full Administrator

- **Important:** The process described here requires Cloudera Manager and cluster downtime.

After you have installed Cloudera Manager and created a cluster, you may need to update the names of the hosts running the Cloudera Manager Server or cluster services. To update a deployment with new hostnames, follow these steps:

1. Verify if SSL/TLS certificates have been issued for any of the services and make sure to create new SSL/TLS certificates in advance for services protected by SSL/TLS. See [Encryption](#).
2. [Export](#) the Cloudera Manager configuration using one of the following methods:
  - Open a browser and go to this URL `http://cm_hostname:7180/api/api_version/cm/deployment`. Save the displayed configuration.

- From terminal type:

```
$ curl -u admin:admin http://cm_hostname:7180/api/api_version/cm/deployment > cme-cm-export.json
```

If Cloudera Manager SSL is in use, specify the `-k` switch:

```
$ curl -k -u admin:admin http://cm_hostname:7180/api/api_version/cm/deployment > cme-cm-export.json
```

where `cm_hostname` is the name of the Cloudera Manager host and `api_version` is the correct [version](#) of the API for the version of Cloudera Manager you are using. For example,  
`http://tcdn5-1.ent.cloudera.com:7180/api/v9/cm/deployment`.

3. [Stop all services](#) on the cluster.
4. [Stop the Cloudera Management Service](#).
5. [Stop the Cloudera Manager Server](#).
6. [Stop the Cloudera Manager Agents](#) on the hosts that will be having the hostname changed.
7. [Back up the Cloudera Manager Server database](#) using `mysqldump`, `pg_dump`, or another preferred backup utility. Store the backup in a safe location.
8. Update names and principals:
  - a. Update the target hosts using standard per-OS/name service methods (`/etc/hosts`, `dns`, `/etc/sysconfig/network`, `hostname`, and so on). Ensure that you remove the old hostname.
  - b. If you are changing the hostname of the host running Cloudera Manager Server do the following:
    - a. Change the hostname per [step 8.a](#).
    - b. Update the Cloudera Manager hostname in `/etc/cloudera-scm-agent/config.ini` on all Agents.
  - c. If the cluster is configured for Kerberos security, do the following:
    - a. In the Cloudera Manager database, set the `merged_keytab` value:
      - **PostgreSQL**

```
update roles set merged_keytab=NULL;
```
      - **MySQL**

```
update ROLES set MERGED_KEYTAB=NULL;
```
    - b. Remove old hostname cluster service principals from the KDC database using one of the following:
      - Use the `delprinc` command within `kadmin.local` interactive shell.
      - From the command line:

```
kadmin.local -q "listprincs" | grep -E "(HTTP|hbase|hdfs|hive|httpfs|hue|impala|mapred|solr|oozie|yarn|zookeeper)[^/]*/*" | sed 's/.*@/cluster-princ.txt/' > cluster-princ.txt
```

Open `cluster-princ.txt` and remove any non-cluster service principal entries within it. Make sure that the default `krbtgt` and other principals you created, or were created by Kerberos by default, are not removed by running the following: `for i in $(cat cluster-princ.txt); do yes yes | kadmin.local -q "delprinc $i"; done`.
    - c. Start the Cloudera Manager database and Cloudera Manager Server.
    - d. Start the Cloudera Manager Agents on the newly renamed hosts. The Agents should show a current heartbeat in Cloudera Manager.
    - e. Within the Cloudera Manager Admin Console recreate all the principals based on the new hostnames:
      - a. Select **Administration > Kerberos**.

- b. Do one of the following:
          - If there are no principals listed, click the **Generate Principals** button.
          - If there are principals listed, click the top checkbox to select all principals and click the **Regenerate** button.
  9. If one of the hosts that was renamed has a NameNode configured with high availability and automatic failover enabled, reconfigure the ZooKeeper Failover Controller znodes to reflect the new hostname.
    - a. Start ZooKeeper Servers.
 

- **Warning:** All other services, and most importantly HDFS, and the ZooKeeper Failover Controller (FC) role within the HDFS, should not be running.
    - b. On one of the hosts that has a ZooKeeper Server role, run `ZooKeeper_HOME/bin/zkCli.sh`, where `ZooKeeper_HOME` is:
      - **Package installation** - `/usr/lib/zookeeper`
      - **Parcel installation** - `Parcel_HOME/CDH/lib/zookeeper`, where `Parcel_HOME` is `/opt/cloudera/parcels` by default.
    - a. If the cluster is configured for Kerberos security, configure ZooKeeper authorization as follows:
      - a. Go to the HDFS service.
      - b. Click the **Instances** tab.
      - c. Click the **Failover Controller** role.
      - d. Click the **Process** tab.
      - e. In the Configuration Files column of the `hdfs/hdfs.sh [ "zkfc" ]` program, expand **Show**.
      - f. Inspect `core-site.xml` in the displayed list of files and determine the value of the `ha.zookeeper.auth` property, which will be something like:  
`digest:hdfs-fcs:TEbW2bgoODa96rO3ZTn7ND5fSOGx0h`. The part after `digest:hdfs-fcs:` is the password (in the example it is `TEbW2bgoODa96rO3ZTn7ND5fSOGx0h`)
      - g. Run the `addauth` command with the password:
 

```
addauth digest hdfs-fcs:TEbW2bgoODa96rO3ZTn7ND5fSOGx0h
```
    - b. Verify that the HA znode exists: `zkCli$ ls /hadoop-ha`.
    - c. Delete the HDFS znode: `zkCli$ rmr /hadoop-ha/nameservicel`.
    - d. If you *are not* running JobTracker in a high availability configuration, delete the HA znode: `zkCli$ rmr /hadoop-ha`.
  - c. In the Cloudera Manager Admin Console, go to the HDFS service.
  - d. Click the **Instances** tab.
  - e. Select **Actions > Initialize High Availability State in ZooKeeper...**
10. Update the Hive metastore:
  - a. Back up the Hive metastore database.
  - b. Go the Hive service.
  - c. Select **Actions > Update Hive Metastore NameNodes** and confirm the command.
11. If you are using the embedded database, update the **Database Hostname** property for each of the Cloudera Management Service roles (Reports Manager, Activity Monitor, Navigator Audit and Metadata Server) and the Hive Metastore Server database hostname.
12. Start all cluster services.
13. Start the Cloudera Management Service.

14. Deploy client configurations.

### Configuring Network Settings

Required Role: **Full Administrator**

To configure a proxy server thorough which data is downloaded to and uploaded from the Cloudera Manager Server, do the following:

1. Select **Administration > Settings**.
2. Click the **Network** category.
3. Configure proxy properties.
4. Click **Save Changes** to commit the changes.


### Managing Alerts

Required Role: **Full Administrator**

The **Administration > Alerts** page provides a summary of the settings for alerts in your clusters.

**Alert Type** The left column lets you select by alert type (Health, Log, or Activity) and within that by service instance. In the case of Health alerts, you can look at alerts for Hosts as well. You can select an individual service to see just the alert settings for that service.

**Health/Log/Activity Alert Settings** Depending on your selection in the left column, the right hand column show you the list of alerts that are enabled or disabled for the selected service type.

To change the alert settings for a service, click the  next to the service name. This will take you to the Monitoring section of the Configuration tab for the service. From here you can enable or disable alerts and configure thresholds as needed.

**Recipients** You can also view the list of recipients configured for the enabled alerts.

#### Configuring Alert Delivery

When you install Cloudera Manager you can configure the mail server you will use with the Alert Publisher. However, if you need to change these settings, you can do so under the Alert Publisher section of the Management Services configuration tab. Under the Alert Publisher role of the Cloudera Manager Management Service, you can configure email or SNMP delivery of alert notifications.


### Configuring Alert Email Delivery

Required Role: **Full Administrator**

#### Sending A Test Alert E-mail

Select the **Administration > Alerts** tab and click the **Send Test Alert** link.

#### Configuring the List Of Alert Recipient Email Addresses

1. Do one of the following:
  - Select the **Administration > Alerts** tab and click the  to the right of **Recipient(s)**.
  - 1. Do one of the following:
    - Select **Clusters > Cloudera Management Service > Cloudera Management Service**.
    - On the Status tab of the Home page, in **Cloudera Management Service** table, click the **Cloudera Management Service** link.
2. Click the **Configuration** tab.



3. Select the **Alert Publisher Default Group** role group.
2. Configure the **Alerts: Mail Message Recipients** property.
3. Click the **Save Changes** button at the top of the page to save your settings.
4. Restart the Alert Publisher role.

### Configuring Alert Email Properties

1. [Display the Cloudera Management Service](#) status page.
2. Click the **Configuration** tab.
3. Select the **Alert Publisher Default Group** role group to see the list of properties. In order to receive email alerts you must set (or verify) the following settings:
  - Enable email alerts
  - Email protocol to use.
  - Your mail server hostname and port.
  - The username and password of the email user that will be logged into the mail server as the "sender" of the alert emails.
  - A comma-separated list of email addresses that will be the recipients of alert emails.
  - The format of the email alert message. Select **json** if you need the message to be parsed by a script or program.
4. Click the **Save Changes** button at the top of the page to save your settings.
5. Restart the Alert Publisher role.

### Configuring Alert SNMP Delivery

[Required Role:](#) **Full Administrator**

- **Important: This feature is available only with a Cloudera Enterprise license.**

For other licenses, the following applies:

- Cloudera Express- The feature is not available.
- Cloudera Enterprise Data Hub Edition Trial - The feature is available until you end the trial or the trial license expires.

To obtain a license for Cloudera Enterprise, fill in this [form](#) or call 866-843-7207. After you install a Cloudera Enterprise license, the feature will be available.

### Enabling, Configuring, and Disabling SNMP Traps

1. Before you enable SNMP traps, configure the trap receiver (Network Management System or SNMP server) with the Cloudera MIB.
2. Do one of the following:
  - Select **Clusters > Cloudera Management Service > Cloudera Management Service**.
  - On the Status tab of the Home page, in **Cloudera Management Service** table, click the **Cloudera Management Service** link.
3. Click the **Configuration** tab.
4. Select **Alert Publisher Default Group > SNMP**.
  - Enter the DNS name or IP address of the Network Management System (SNMP server) acting as the trap receiver in the SNMP NMS Hostname property.

## Cloudera Manager Administration

- In the SNMP Security Level property, select the version of SNMP you are using: SNMPv2, SNMPv3 without authentication and without privacy (`noAuthNoPriv`), or SNMPv3 with authentication and without privacy (`authNoPriv`) and specify the required properties:
  - SNMPv2 – SNMPv2 Community String.
  - SNMPv3 without authentication (`noAuthNoPriv`) – SNMP Server Engine Id and SNMP Security UserName.
  - SNMPv3 with authentication (`authNoPriv`) – SNMP Server Engine Id, SNMP Security UserName, SNMP Authentication Protocol, and SNMP Authentication Protocol Pass Phrase.
- You can also change other settings such as the port, retry, or timeout values.

5. Click **Save Changes** when you are done.

6. Restart the Alert Publisher role.

To disable SNMP traps, remove the hostname from the **SNMP NMS Hostname** property (`alert.snmp.server.hostname`).

### Viewing the Cloudera MIB

1. Do one of the following:
  - Select **Clusters > Cloudera Management Service > Cloudera Management Service**.
  - On the Status tab of the Home page, in **Cloudera Management Service** table, click the **Cloudera Management Service** link.
2. Click the **Configuration** tab.
3. Select **Alert Publisher Default Group > SNMP**.
4. In the **Description** column for the first property (**SNMP NMS Hostname**) click the **SNMP Mib** link.

## Managing Licenses

**Required Role:** Full Administrator

When you install Cloudera Manager, you can choose to select Cloudera Express(no license required), a 60-day Cloudera Enterprise Data Hub Edition trial license, or Cloudera Enterprise(which requires a license). You can later end a trial license or upgrade your license.

### About Trial Licenses

You can use the trial license only once; once the 60-day trial period has expired or you have ended the trial, you cannot restart it.


When a trial ends, features that require a Cloudera Enterprise license immediately become unavailable. However, data or configurations associated with the disabled functions are not deleted, and become available again when you install a Cloudera Enterprise license. Trial expiration or termination has the following effects:

- Only local users can log in (no LDAP or SAML authentication).
- Configuration history is unavailable.
- Alerts cannot be delivered as SNMP traps.
- Operational reports are inaccessible (but remain in the database).
- Commands such as Rolling Restart, History and Rollback (under the Configuration tab), Send Diagnostic Data, and starting Cloudera Navigator roles are disabled or not available.

### Accessing the License Page

To access the license page, select **Administration > License**.

If you have a license installed, the license page indicates its status (for example, whether your license is currently valid) and displays the license details: the license owner, the license key, and the expiration date of the license, if there is one.

At the right side of the page a table shows the usage of licensed components based on the number of hosts with those products installed. You can move the cursor over the  to see an explanation of each item.

- **Basic Edition** - a cluster running core CDH services: HDFS, Hive, Hue, MapReduce, Oozie, Sqoop, YARN, and ZooKeeper.
- **Flex Edition** - a cluster running core CDH services plus one of the following: Accumulo, HBase, Impala, Navigator, Solr, Spark.
- **Data Hub Edition** - a cluster running core CDH services plus any of the following: Accumulo, HBase, Impala, Navigator, Solr, Spark.

### Ending a Cloudera Enterprise Data Hub Edition Trial

If you are using the trial edition the License page indicates when your license will expire. However, you can end the trial at any time (prior to expiration) as follows:

1. On the License page, click **End Trial**.
2. Confirm that you want to end the trial.
3. Restart the Cloudera Management Service, HBase, HDFS, and Hive services to pick up configuration changes.

### Upgrading from Cloudera Express to a Cloudera Enterprise Data Hub Edition Trial

To start a trial, on the License page, click **Try Cloudera Enterprise Data Hub Edition for 60 Days**.

1. Cloudera Manager displays a pop-up describing the features enabled with Cloudera Enterprise Data Hub Edition. Click **OK** to proceed. At this point, your installation is upgraded and the Customize Role Assignments page displays.
2. Under **Reports Manager** click **Select a host**. The pageable host selection dialog displays.

The following shortcuts for specifying hostname patterns are supported:

- Range of hostnames (without the domain portion)

Range Definition	Matching Hosts
10.1.1.[1-4]	10.1.1.1, 10.1.1.2, 10.1.1.3, 10.1.1.4
host[1-3].company.com	host1.company.com, host2.company.com, host3.company.com
host[07-10].company.com	host07.company.com, host08.company.com, host09.company.com, host10.company.com

- IP addresses
  - Rack name
3. Select a host and click **OK**.
  4. When you are satisfied with the assignments, click **Continue**. The Database Setup screen displays.
  5. Configure database settings:
    - a. Choose the database type:
      - Leave the default setting of **Use Embedded Database** to have Cloudera Manager create and configure required databases. Make a note of the auto-generated passwords.

## Cluster Setup

### Database Setup

Configure and test database connections. If using custom databases, create the databases first according to the [Installing and Configuring an External Database](#) section of the [Installation Guide](#) .

☐ Use Custom Databases  
☒ Use Embedded Database

When using the embedded database, passwords are automatically generated. Please copy them down.

Service	Database Host Name	Database Type	Database Name	Username	Password
<b>Hive</b>	tcdn53-1.ent.cloudera.com:7432	PostgreSQL	hive	hive	24FLyrj0zb
<b>Activity Monitor</b>	tcdn53-1.ent.cloudera.com:7432	PostgreSQL	amon	amon	2VCic0tDJE
<b>Reports Manager</b>	tcdn53-1.ent.cloudera.com:7432	PostgreSQL	rman	rman	Mn2l8tEoCH
<b>Navigator Audit Server</b>	tcdn53-1.ent.cloudera.com:7432	PostgreSQL	nav	nav	P89RAR6e0o
<b>Navigator Metadata Server</b>	tcdn53-1.ent.cloudera.com:7432	PostgreSQL	navms	navms	29O536GxZp

- Select **Use Custom Databases** to specify external databases.

1. Enter the database host, database type, database name, username, and password for the database that you created when you set up the database.
- b. Click **Test Connection** to confirm that Cloudera Manager can communicate with the database using the information you have supplied. If the test succeeds in all cases, click **Continue**; otherwise check and correct the information you have provided for the database and then try the test again. (For some servers, if you are using the embedded database, you will see a message saying the database will be created at a later step in the installation process.) The Review Changes screen displays.
6. Review the configuration changes to be applied. Confirm the settings entered for file system paths. The file paths required vary based on the services to be installed.

▪ **Warning:** DataNode data directories should not be placed on NAS devices.

Click **Continue**. The wizard starts the services.

7. At this point, your installation is upgraded. Click **Continue**.
8. Restart Cloudera Management Services and audited services to pick up configuration changes. The audited services will write audit events to a log file, but the events are not transferred to the Cloudera Navigator Audit Server until you add and start the Cloudera Navigator Audit Server role as described in [Adding Cloudera Navigator Roles](#) on page 317. For information on Cloudera Navigator, see [Cloudera Navigator documentation](#).

## Upgrading from a Cloudera Enterprise Data Hub Edition Trial to Cloudera Enterprise

1. Purchase a Cloudera Enterprise license from Cloudera.
2. On the License page, click **Upload License**.
3. Click the document icon to the left of the **Select a License File** text field.
4. Navigate to the location of your license file, click the file, and click **Open**.
5. Click **Upload**.

### Upgrading from Cloudera Express to Cloudera Enterprise

1. Purchase a Cloudera Enterprise license from Cloudera.
2. On the License page, click **Upload License**.
3. Click the document icon to the left of the **Select a License File** text field.
4. Navigate to the location of your license file, click the file, and click **Open**.
5. Click **Upload**.
6. Cloudera Manager displays a pop-up describing the features enabled with Cloudera Enterprise Data Hub Edition. Click **OK** to proceed. At this point, your installation is upgraded and the Customize Role Assignments page displays.
7. Under **Reports Manager** click **Select a host**. The pageable host selection dialog displays.

The following shortcuts for specifying hostname patterns are supported:

- Range of hostnames (without the domain portion)

Range Definition	Matching Hosts
10.1.1.[1-4]	10.1.1.1, 10.1.1.2, 10.1.1.3, 10.1.1.4
host[1-3].company.com	host1.company.com, host2.company.com, host3.company.com
host[07-10].company.com	host07.company.com, host08.company.com, host09.company.com, host10.company.com

- IP addresses
  - Rack name
8. When you are satisfied with the assignments, click **Continue**. The Database Setup screen displays.
  9. Configure database settings:
    - a. Choose the database type:
      - Leave the default setting of **Use Embedded Database** to have Cloudera Manager create and configure required databases. Make a note of the auto-generated passwords.

## Cluster Setup

### Database Setup

Configure and test database connections. If using custom databases, create the databases first according to the [Installing and Configuring an External Database](#) section of the [Installation Guide](#) .

☐ Use Custom Databases  
☒ Use Embedded Database

When using the embedded database, passwords are automatically generated. Please copy them down.

Service	Database Host Name	Database Type	Database Name	Username	Password
<b>Hive</b>	tcdn53-1.ent.cloudera.com:7432	PostgreSQL	hive	hive	24FLYrj0zb
<b>Activity Monitor</b>	tcdn53-1.ent.cloudera.com:7432	PostgreSQL	amon	amon	2VCic0tDJE
<b>Reports Manager</b>	tcdn53-1.ent.cloudera.com:7432	PostgreSQL	rman	rman	Mn2l8tEoCH
<b>Navigator Audit Server</b>	tcdn53-1.ent.cloudera.com:7432	PostgreSQL	nav	nav	P89FAR6e0o
<b>Navigator Metadata Server</b>	tcdn53-1.ent.cloudera.com:7432	PostgreSQL	navms	navms	29O536GxZp

- Select **Use Custom Databases** to specify external databases.

1. Enter the database host, database type, database name, username, and password for the database that you created when you set up the database.

- b. Click **Test Connection** to confirm that Cloudera Manager can communicate with the database using the information you have supplied. If the test succeeds in all cases, click **Continue**; otherwise check and correct the information you have provided for the database and then try the test again. (For some servers, if you are using the embedded database, you will see a message saying the database will be created at a later step in the installation process.) The Review Changes screen displays.

10. Review the configuration changes to be applied. Confirm the settings entered for file system paths. The file paths required vary based on the services to be installed.

- **Warning:** DataNode data directories should not be placed on NAS devices.

Click **Continue**. The wizard starts the services.

11. At this point, your installation is upgraded. Click **Continue**.
12. Restart Cloudera Management Services and audited services to pick up configuration changes. The audited services will write audit events to a log file, but the events are not transferred to the Cloudera Navigator Audit Server until you add and start the Cloudera Navigator Audit Server role as described in [Adding Cloudera Navigator Roles](#) on page 317. For information on Cloudera Navigator, see [Cloudera Navigator documentation](#).

If you want to use the Cloudera Navigator Metadata Server, add its role following the instructions in [Adding Cloudera Navigator Roles](#) on page 317.

## Renewing a License

1. Download the license file and save it locally.
2. In Cloudera Manager, go to the **Home** page.
3. Select **Administration > License**.

4. Click **Upload License**.
5. Browse to the license file you downloaded.
6. Click **Upload**.

You do not need to restart Cloudera Manager for the new license to take effect.

## Sending Usage and Diagnostic Data to Cloudera

**Required Role:** **Full Administrator**

Cloudera Manager collects anonymous usage information and takes regularly-scheduled snapshots of the state of your cluster and automatically sends them anonymously to Cloudera. This helps Cloudera improve and optimize Cloudera Manager.

If you have a Cloudera Enterprise license, you can also trigger the collection of diagnostic data and send it to Cloudera Support to aid in resolving a problem you may be having.

### Configuring a Proxy Server

To configure a proxy server through which usage and diagnostic data is uploaded, follow the instructions in [Configuring Network Settings](#) on page 304.

### Managing Anonymous Usage Data Collection

Cloudera Manager sends anonymous usage information using Google Analytics to Cloudera. The information helps Cloudera improve Cloudera Manager. By default anonymous usage data collection is *enabled*.

1. Select **Administration > Settings**.
2. Under the **Other** category, set the **Allow Usage Data Collection** property.
3. Click **Save Changes** to commit the changes.

### Managing Hue Analytics Data Collection

**Required Role:** **Configurator** **Cluster Administrator** **Full Administrator**

Hue tracks anonymised pages and application versions in order to gather information to help compare each application's usage levels. The data collected does not include any hostnames or IDs. For example, the data is of the form: /2.3.0/pig, /2.5.0/beeswax/execute. You can restrict data collection as follows:

1. Go to the Hue service.
2. Click the **Configuration** tab.
3. Expand the **Service-Wide** category.
4. Uncheck the **Enable Usage Data Collection** checkbox.
5. Click **Save Changes** to commit the changes.
6. Restart the Hue service.

### Diagnostic Data Collection

To help with solving problems when using Cloudera Manager on your cluster, Cloudera Manager collects diagnostic data on a regular schedule, and automatically sends it to Cloudera. By default Cloudera Manager is configured to collect data weekly and to send it *automatically*. You can schedule the frequency of data collection on a daily, weekly, or monthly schedule, or disable the scheduled collection of data entirely. You can also send a collected data set [manually](#).

■ **Note:**

- Automatically sending diagnostic data requires the Cloudera Manager Server host to have Internet access, and be configured for sending data automatically. If your Cloudera Manager server does not have Internet access, and you have a Cloudera Enterprise license, you can manually send the diagnostic data as described in [Manually Triggering Collection and Transfer of Diagnostic Data to Cloudera](#) on page 313.
- Automatically sending diagnostic data may fail sometimes and return an error message of "Could not send data to Cloudera." To work around this issue, you can manually send the data to Cloudera Support.

### What Data Does Cloudera Manager Collect?

Cloudera Manager collects and returns a significant amount of information about the health and performance of the cluster. It includes:

- Up to 1000 Cloudera Manager audit events: Configuration changes, add/remove of users, roles, services, etc.
- One day's worth of Cloudera Manager events: This includes critical errors Cloudera Manager watches for and more
- Data about the cluster structure which includes a list of all hosts, roles, and services along with the configurations that are set through Cloudera Manager. Where passwords are set in Cloudera Manager, the passwords are not returned.
- Cloudera Manager license and version number.
- Current health information for hosts, service, and roles. Includes results of health tests run by Cloudera Manager.
- Heartbeat information from each host, service, and role. These include status and some information about memory, disk, and processor usage.
- The results of running Host Inspector.
- One day's worth of Cloudera Manager metrics.

■ **Note:** If you are using Cloudera Express, host metrics are not included.

- A download of the debug pages for Cloudera Manager roles.
- For each host in the cluster, the result of running a number of system-level commands on that host.
- Logs from each role on the cluster, as well as the Cloudera Manager server and agent logs.
- Which parcels are activated for which clusters.
- Whether there's an active trial, and if so, metadata about the trial.
- Metadata about the Cloudera Manager server, such as its JMX metrics, stack traces, and the database/host it's running with.
- HDFS/Hive replication schedules (including command history) for the deployment.
- Impala query logs.

### Configuring the Frequency of Diagnostic Data Collection

By default, Cloudera Manager collects diagnostic data on a weekly basis. You can change the frequency to daily, weekly, monthly, or never. If you are a Cloudera Enterprise customer and you set the schedule to **never** you can still collect and send data to Cloudera on demand. If you are a Cloudera Express customer and you set the schedule to **never**, data is not collected or sent to Cloudera.

1. Select **Administration > Settings**.
2. Under the **Support** category, click **Scheduled Diagnostic Data Collection Frequency** and select the frequency.
3. To set the day and time of day that the collection will be performed, click **Scheduled Diagnostic Data Collection Time** and specify the date and time in the pop-up control.
4. Click **Save Changes** to commit the changes.



You can see the current setting of the data collection frequency by viewing **Support > Scheduled Diagnostics**: in the main navigation bar.

### Specifying the Diagnostic Data Directory

You can configure the directory where collected data is stored.

1. Select **Administration > Settings**.
2. Under the **Support** category, set the **Diagnostic Data Bundle Directory** to a directory on the host running Cloudera Manager Server. The directory must exist and be enabled for writing by the user `cloudera-scm`. If this field is left blank, the data is stored in `/tmp`.
3. Click **Save Changes** to commit the changes.

### Collecting and Sending Diagnostic Data to Cloudera

- **Important:** This feature is available only with a Cloudera Enterprise license.

For other licenses, the following applies:

- Cloudera Express- The feature is not available.
- Cloudera Enterprise Data Hub Edition Trial - The feature is available until you end the trial or the trial license expires.

To obtain a license for Cloudera Enterprise, fill in this [form](#) or call 866-843-7207. After you install a Cloudera Enterprise license, the feature will be available.

### Disabling the Automatic Sending of Diagnostic Data from a Manually Triggered Collection

If you do not want data automatically sent to Cloudera after manually triggering data collection, you can disable this feature. The data you collect will be saved and can be downloaded for sending to Cloudera Support at a later time.

1. Select **Administration > Settings**.
2. Under the **Support** category, uncheck the box for **Send Diagnostic Data to Cloudera Automatically**.
3. Click **Save Changes** to commit the changes.

- **Note:** The Send Diagnostic Data form that displays when you collect data in one of the following procedures indicates whether the data will be sent automatically.

### Manually Triggering Collection and Transfer of Diagnostic Data to Cloudera

1. Optionally change the System Identifier property:
  - a. Select **Administration > Settings**.
  - b. Under the **Other** category, set the System Identifier property and click **Save Changes**.
2. Under the **Support** menu at the top right of the navigation bar, choose **Send Diagnostic Data**. The Send Diagnostic Data form displays.
3. Fill in or change the information here as appropriate:
  - Cloudera Manager populates the **End Time** based on the setting of the Time Range selector. You should change this to be a few minutes after you observed the problem or condition that you are trying to capture. The time range is based on the timezone of the host where Cloudera Manager Server is running.
  - If you have a support ticket open with Cloudera Support, include the support ticket number in the field provided.
4. Depending on whether you have disabled automatic sending of data, do one of the following:

- Click **Collect and Send Diagnostic Data**. A Running Commands window shows you the progress of the data collection steps. When these steps are complete, the collected data is sent to Cloudera.
- Click **Collect Diagnostic Data**. A Command Details window shows you the progress of the data collection steps.
  1. In the Command Details window, click **Download Result Data** to download and save a zip file of the information.
  2. Send the data to Cloudera Support by doing one of the following:
    - Send the bundle using a Python script:
      1. Download the [phone\\_home](#) script.
      2. Copy the script and the downloaded data file to a host that has Internet access.
      3. Run the following command on that host:

```
python phone_home.py --file downloaded_data_file
```

- Attach the bundle to the SFDC case. Do not rename the bundle as this can cause a delay in processing the bundle.
- Contact [Cloudera Support](#) and arrange to send the data file.

## Exporting and Importing Cloudera Manager Configuration

You can use the Cloudera Manager API to programmatically export and import a definition of all the entities in your Cloudera Manager-managed deployment—clusters, service, roles, hosts, users and so on. See the [Cloudera Manager API](#) documentation on how to manage deployments using the [/cm/deployment](#) resource.

## Other Cloudera Manager Tasks and Settings

From the **Administration** tab you can select options for configuring settings that affect how Cloudera Manager interacts with your clusters.

### Settings

The **Settings** page provides a number of categories as follows:

- **Performance** - Set the Cloudera Manager Agent heartbeat interval. See [Configuring Agent Heartbeat and Health Status Options](#) on page 297.
- **Advanced** - Enable API debugging and other advanced options.
- **Monitoring** - Set Agent health status parameters. For configuration instructions, see [Configuring Cloudera Manager Agents](#) on page 297.
- **Security** - Set TLS encryption settings to enable TLS encryption between the Cloudera Manager Server, Agents, and clients. For configuration instructions, see [Configuring TLS Security for Cloudera Manager](#). You can also:
  - Set the realm for Kerberos security and point to a custom keytab retrieval script. For configuration instructions, see [Cloudera Security](#).
  - Specify session timeout and a "Remember Me" option.
- **Ports and Addresses** - Set ports for the Cloudera Manager Admin Console and Server. For configuration instructions, see [Configuring Cloudera Manager Server Ports](#) on page 295.
- **Other**
  - Enable Cloudera usage data collection For configuration instructions, see [Managing Anonymous Usage Data Collection](#) on page 311.
  - Set a custom header color and banner text for the Admin console.

- Set an "Information Assurance Policy" statement – this statement will be presented to every user before they are allowed to access the login dialog. The user must click "I Agree" in order to proceed to the login dialog.
- Disable/enable the auto-search for the Events panel at the bottom of a page.
- **Support**
  - Configure diagnostic data collection properties. See [Diagnostic Data Collection](#) on page 311.
  - Configure how to access Cloudera Manager [help](#) files.
- **External Authentication** – Specify the configuration to use LDAP, Active Directory, or an external program for authentication. See [Configuring External Authentication for Cloudera Manager](#) for instructions.
- **Parcels** – Configure settings for parcels, including the location of remote repositories that should be made available for download, and other settings such as the frequency with which Cloudera Manager will check for new parcels, limits on the number of downloads or concurrent distribution uploads. See [Parcels](#) for more information.
- **Network** – Configure proxy server settings. See [Configuring Network Settings](#) on page 304.
- **Custom Service Descriptors** – Configure custom service descriptor properties for [Add-on Services](#) on page 31.

## Alerts

See [Managing Alerts](#) on page 304.

## Users

See [Cloudera Manager User Accounts](#).

## Kerberos

See [Enabling Kerberos Authentication Using the Wizard](#).

## License

See [Managing Licenses](#) on page 306.

## User Interface Language

You can change the language of the Cloudera Manager Admin Console User Interface through the language preference in your browser. Information on how to do this for the browsers supported by Cloudera Manager is shown under the Administration page. You can also change the language for the information provided with activity and health events, and for alert email messages by selecting **Language**, selecting the language you want from the drop-down list on this page, then clicking **Save Changes**.

## Peers

See [Designating a Replication Source](#) on page 268.

## Displaying the Cloudera Manager Server Version and Server Time

To display the version, build number, and time for the Cloudera Manager Server:

1. Open the Cloudera Manager Admin Console.
2. Select **Support > About**.

## Displaying Cloudera Manager Documentation

To display Cloudera Manager documentation:

1. Open the Cloudera Manager Admin Console.

2. Select **Support > Help, Installation Guide, API Documentation, or Release Notes**. By default, the Help and Installation Guide files from the Cloudera web site are opened. This is because local help files are not updated after installation. You can configure Cloudera Manager to open either the latest Help and Installation Guide from the Cloudera web site (this option requires Internet access from the browser) or locally-installed Help and Installation Guide by configuring the **Administration > Settings > Support > Open latest Help files from the Cloudera website** property.

## Cloudera Management Service

The Cloudera Management Service implements various management features as a set of roles:

- Activity Monitor - collects information about activities run by the MapReduce service. This role is not added by default.
- Host Monitor - collects health and metric information about hosts
- Service Monitor - collects health and metric information about services and activity information from the YARN and Impala services
- Event Server - aggregates relevant Hadoop events and makes them available for alerting and searching
- Alert Publisher - generates and delivers alerts for certain types of events
- Reports Manager - generates reports that provide an historical view into disk utilization by user, user group, and directory, processing activities by user and YARN pool, and HBase tables and namespaces. This role is not added in Cloudera Express.

Cloudera Manager manages each role separately, instead of as part of the Cloudera Manager Server, for scalability (for example, on large deployments it's useful to put the monitor roles on their own hosts) and isolation.


In addition, for certain editions of the Cloudera Enterprise license, the Cloudera Management Service provides the [Navigator Audit Server](#) and [Navigator Metadata Server](#) roles for [Cloudera Navigator](#).

### Displaying the Cloudera Management Service Status

1. Do one of the following:
  - Select **Clusters > Cloudera Management Service > Cloudera Management Service**.
  - On the Status tab of the Home page, in **Cloudera Management Service** table, click the **Cloudera Management Service** link.

### Starting the Cloudera Management Service

**Required Role:** **Cluster Administrator** **Full Administrator**

1. Do one of the following:
  - 1. Select **Clusters > Cloudera Management Service > Cloudera Management Service**.
  - 2. Select **Actions > Start**.
  - 1. On the Home page, click  to the right of **Cloudera Management Service** and select **Start**.
2. Click **Start** to confirm. The **Command Details** window shows the progress of starting the roles.
3. When **Command completed with n/n successful subcommands** appears, the task is complete. Click **Close**.

### Stopping the Cloudera Management Service

**Required Role:** **Cluster Administrator** **Full Administrator**

1. Do one of the following:
  - 1. Select **Clusters > Cloudera Management Service > Cloudera Management Service**.

2. Select **Actions > Stop**.

- 1. On the Home page, click  to the right of **Cloudera Management Service** and select **Stop**.

2. Click **Stop** to confirm. The **Command Details** window shows the progress of stopping the roles.

3. When **Command completed with n/n successful subcommands** appears, the task is complete. Click **Close**.

### Restarting the Cloudera Management Service

**Required Role:** Cluster Administrator Full Administrator

1. Do one of the following:

- 1. Select **Clusters > Cloudera Management Service > Cloudera Management Service**.
- 2. Select **Actions > Restart**.

- On the Home page, click  to the right of **Cloudera Management Service** and select **Restart**.

2. Click **Restart** to confirm. The **Command Details** window shows the progress of stopping and then starting the roles.

3. When **Command completed with n/n successful subcommands** appears, the task is complete. Click **Close**.

### Configuring Management Service Database Limits

**Required Role:** Cluster Administrator Full Administrator

Each Cloudera Management Service role maintains a database for retaining the data it monitors. These databases (as well as the log files maintained by these services) can grow quite large. For example, the Activity Monitor maintains data at the service level, the activity level (MapReduce jobs and aggregate activities), and at the task attempt level. Limits on these data sets are configured when you create the management services, but you can modify these parameters through the Configuration settings in the Cloudera Manager Admin Console. For example, the Event Server lets you set a total number of events to store, and Activity Monitor gives you "purge" settings (also in hours) for the data it stores.

There are also settings for the logs that these various services create. You can throttle how big the logs are allowed to get and how many previous logs to retain.

1. Do one of the following:

- Select **Clusters > Cloudera Management Service > Cloudera Management Service**.
- On the Status tab of the Home page, in **Cloudera Management Service** table, click the **Cloudera Management Service** link.

2. Click the **Configuration** tab.

3. In the left-hand column, select the Default role group for the role whose configurations you want to modify.

4. Edit the appropriate properties:

- **Activity Monitor** - the **Purge** or **Expiration** period properties are found in the top-level settings for the role.
- **Host and Service Monitor** - see [Data Storage for Monitoring Data](#).
- **Log Files** - log file size settings will be under the **Logs** category under the role group.

5. Click **Save Changes**.

### Adding Cloudera Navigator Roles

**Required Role:** Navigator Administrator Full Administrator

1. Do one of the following:

- Select **Clusters > Cloudera Management Service > Cloudera Management Service**.
  - On the Status tab of the Home page, in **Cloudera Management Service** table, click the **Cloudera Management Service** link.
2. Click the **Instances** tab.
  3. Click the **Add Role Instances** button. The Customize Role Assignments page displays.
  4. Assign the Navigator role to a host.
    - a. Customize the assignment of role instances to hosts. The wizard evaluates the hardware configurations of the hosts to determine the best hosts for each role. The wizard assigns all worker roles to the same set of hosts to which the HDFS DataNode role is assigned. These assignments are typically acceptable, but you can reassign them if necessary.

Click a field below a role to display a dialog containing a list of hosts. If you click a field containing multiple hosts, you can also select **All Hosts** to assign the role to all hosts or **Custom** to display the pageable hosts dialog.

The following shortcuts for specifying hostname patterns are supported:

- Range of hostnames (without the domain portion)

Range Definition	Matching Hosts
10.1.1.[1-4]	10.1.1.1, 10.1.1.2, 10.1.1.3, 10.1.1.4
host[1-3].company.com	host1.company.com, host2.company.com, host3.company.com
host[07-10].company.com	host07.company.com, host08.company.com, host09.company.com, host10.company.com

- IP addresses
- Rack name

Click the **View By Host** button for an overview of the role assignment by hostname ranges.

5. When you are satisfied with the assignments, click **Continue**. The Database Setup screen displays.
6. Configure database settings:
  - a. Choose the database type:
    - Leave the default setting of **Use Embedded Database** to have Cloudera Manager create and configure required databases. Make a note of the auto-generated passwords.

## Cluster Setup

### Database Setup

Configure and test database connections. If using custom databases, create the databases first according to the [Installing and Configuring an External Database](#) section of the [Installation Guide](#) .

☐ Use Custom Databases  
☒ Use Embedded Database

When using the embedded database, passwords are automatically generated. Please copy them down.

Database Name	Database Type	Database Name	Username	Password
Hive	PostgreSQL	hive	hive	24FLYrj0zb
Activity Monitor	PostgreSQL	amon	amon	2VCic0tDJE
Reports Manager	PostgreSQL	rman	rman	Mn2l8tEoCH
Navigator Audit Server	PostgreSQL	nav	nav	P89RAR6e0o
Navigator Metadata Server	PostgreSQL	navms	navms	29O536GxZp

- Select **Use Custom Databases** to specify external databases.
  1. Enter the database host, database type, database name, username, and password for the database that you created when you set up the database.
  - b. Click **Test Connection** to confirm that Cloudera Manager can communicate with the database using the information you have supplied. If the test succeeds in all cases, click **Continue**; otherwise check and correct the information you have provided for the database and then try the test again. (For some servers, if you are using the embedded database, you will see a message saying the database will be created at a later step in the installation process.) The Review Changes screen displays.

7. Click **Finish**.

## Deleting Cloudera Navigator Roles

**Required Role:** **Navigator Administrator** **Full Administrator**

1. Do one of the following:
  - Select **Clusters > Cloudera Management Service > Cloudera Management Service**.
  - On the Status tab of the Home page, in **Cloudera Management Service** table, click the **Cloudera Management Service** link.
2. Click the **Instances** tab.
3. Check the checkboxes next to the **Navigator Audit Server** and **Navigator Metadata Server** roles.
4. Do one of the following depending on your role:
  - **Required Role:** **Full Administrator**
    1. Check the checkboxes next to the **Navigator Audit Server** and **Navigator Metadata Server** roles.
    2. Select **Actions for Selected > Stop** and click **Stop** to confirm.

- **Required Role:** **Navigator Administrator** **Full Administrator**
  1. Click the **Navigator Audit Server** role link.
  2. Select **Actions** > **Stop this Navigator Audit Server** and click **Stop this Navigator Audit Server** to confirm.
  3. Click the **Navigator Metadata Server** role link.
  4. Select **Actions** > **Stop this Navigator Metadata Server** and click **Stop this Navigator Metadata Server** to confirm.
- 5. Check the checkboxes next to the **Navigator Audit Server** and **Navigator Metadata Server** roles.
- 6. Select **Actions for Selected** > **Delete**. Click **Delete** to confirm the deletion.



# Cloudera Navigator Administration

Cloudera Navigator is implemented as two roles within the [Cloudera Management Service](#) on page 316.

For information on managing the Cloudera Navigator roles, see the following topics.

## Related Information

- [Cloudera Navigator 2 Overview](#)
- [Installing Cloudera Navigator](#)
- [Upgrading Cloudera Navigator](#)
- [Cloudera Data Management](#)
- [Configuring Authentication in Cloudera Navigator](#)
- [Configuring SSL for Cloudera Navigator](#)
- [Cloudera Navigator User Roles](#)

## Cloudera Navigator Audit Server

Describes how to add and configure the Navigator Audit Server role.

- **Important: This feature is available only with a Cloudera Enterprise license.**

For other licenses, the following applies:

- Cloudera Express- The feature is not available.
- Cloudera Enterprise Data Hub Edition Trial - The feature is available until you end the trial or the trial license expires.

To obtain a license for Cloudera Enterprise, fill in this [form](#) or call 866-843-7207. After you install a Cloudera Enterprise license, the feature will be available.

## Adding the Navigator Audit Server Role

**Required Role:** Navigator Administrator Full Administrator

Before adding the Navigator Audit Server role, configure [the database](#) where audit events are stored.

1. Do one of the following:
  - Select **Clusters > Cloudera Management Service > Cloudera Management Service**.
  - On the Status tab of the Home page, in **Cloudera Management Service** table, click the **Cloudera Management Service** link.
2. Click the **Instances** tab.
3. Click the **Add Role Instances** button. The Customize Role Assignments page displays.
4. Assign the Navigator role to a host.
  - a. Customize the assignment of role instances to hosts. The wizard evaluates the hardware configurations of the hosts to determine the best hosts for each role. The wizard assigns all worker roles to the same set of hosts to which the HDFS DataNode role is assigned. These assignments are typically acceptable, but you can reassign them if necessary.

Click a field below a role to display a dialog containing a list of hosts. If you click a field containing multiple hosts, you can also select **All Hosts** to assign the role to all hosts or **Custom** to display the pageable hosts dialog.

The following shortcuts for specifying hostname patterns are supported:

- Range of hostnames (without the domain portion)

Range Definition	Matching Hosts
10.1.1.[1-4]	10.1.1.1, 10.1.1.2, 10.1.1.3, 10.1.1.4
host[1-3].company.com	host1.company.com, host2.company.com, host3.company.com
host[07-10].company.com	host07.company.com, host08.company.com, host09.company.com, host10.company.com

- IP addresses
- Rack name

Click the **View By Host** button for an overview of the role assignment by hostname ranges.

5. When you are satisfied with the assignments, click **Continue**. The Database Setup screen displays.

6. Configure database settings:

a. Choose the database type:

- Leave the default setting of **Use Embedded Database** to have Cloudera Manager create and configure required databases. Make a note of the auto-generated passwords.

Cluster Setup

Database Setup

Configure and test database connections. If using custom databases, create the databases first according to the [Installing and Configuring an External Database](#) section of the [Installation Guide](#) .

☐ Use Custom Databases  
☒ Use Embedded Database

When using the embedded database, passwords are automatically generated. Please copy them down.

**Hive**

Currently assigned to run on `tcdn53-1.ent.cloudera.com`.

Database Host Name:	Database Type:	Database Name :	Username:	Password:
<input type="text" value="tcdn53-1.ent.cloudera.com:7432"/>	<input type="text" value="PostgreSQL"/>	<input type="text" value="hive"/>	<input type="text" value="hive"/>	<input type="text" value="24FLyrjozb"/>

**Activity Monitor**

Currently assigned to run on `tcdn53-1.ent.cloudera.com`.

Database Host Name:	Database Type:	Database Name :	Username:	Password:
<input type="text" value="tcdn53-1.ent.cloudera.com:7432"/>	<input type="text" value="PostgreSQL"/>	<input type="text" value="amon"/>	<input type="text" value="amon"/>	<input type="text" value="2VCic0tDJE"/>

**Reports Manager**

Currently assigned to run on `tcdn53-1.ent.cloudera.com`.

Database Host Name:	Database Type:	Database Name :	Username:	Password:
<input type="text" value="tcdn53-1.ent.cloudera.com:7432"/>	<input type="text" value="PostgreSQL"/>	<input type="text" value="rman"/>	<input type="text" value="rman"/>	<input type="text" value="Mn2l8tEoCH"/>

**Navigator Audit Server**

Currently assigned to run on `tcdn53-1.ent.cloudera.com`.

Database Host Name:	Database Type:	Database Name :	Username:	Password:
<input type="text" value="tcdn53-1.ent.cloudera.com:7432"/>	<input type="text" value="PostgreSQL"/>	<input type="text" value="nav"/>	<input type="text" value="nav"/>	<input type="text" value="P89RAR6e0o"/>

**Navigator Metadata Server**

Currently assigned to run on `tcdn53-1.ent.cloudera.com`.

Database Host Name:	Database Type:	Database Name :	Username:	Password:
<input type="text" value="tcdn53-1.ent.cloudera.com:7432"/>	<input type="text" value="PostgreSQL"/>	<input type="text" value="navms"/>	<input type="text" value="navms"/>	<input type="text" value="29O536GxZp"/>

- Select **Use Custom Databases** to specify external databases.

1. Enter the database host, database type, database name, username, and password for the database that you created when you set up the database.

b. Click **Test Connection** to confirm that Cloudera Manager can communicate with the database using the information you have supplied. If the test succeeds in all cases, click **Continue**; otherwise check and correct the information you have provided for the database and then try the test again. (For some servers, if you

are using the embedded database, you will see a message saying the database will be created at a later step in the installation process.) The Review Changes screen displays.

7. Click **Finish**.

### Starting, Stopping, and Restarting the Navigator Audit Server

1. Do one of the following:

- Select **Clusters > Cloudera Management Service > Cloudera Management Service**.
- On the Status tab of the Home page, in **Cloudera Management Service** table, click the **Cloudera Management Service** link.

2. Click the **Instances** tab.

3. Do one of the following depending on your role:

- **Required Role:** **Full Administrator**
  1. Check the checkbox next to the **Navigator Audit Server** role.
  2. Select **Actions for Selected > Action**. Click *Action* to confirm the action, where *Action* is Start, Stop, or Restart.
- **Required Role:** **Navigator Administrator** **Full Administrator**
  1. Click the **Navigator Audit Server** role link.
  2. Select **Actions > Action this Navigator Audit Server**. Click *Action this Navigator Audit Server*, where *Action* is Start, Stop, or Restart, to confirm the action.

### Configuring the Navigator Audit Server Log Directory

**Required Role:** **Navigator Administrator** **Full Administrator**

1. Do one of the following:

- Select **Clusters > Cloudera Management Service > Cloudera Management Service**.
- On the Status tab of the Home page, in **Cloudera Management Service** table, click the **Cloudera Management Service** link.

2. Click the **Configuration** tab.

3. Expand the **Navigator Audit Server Default Group** category.

4. Set the **Navigator Audit Server Log Directory** property.

5. Click **Save Changes**.

6. Click the **Instances** tab.

7. Check the checkbox next to the **Navigator Audit Server** role.

8. Select **Actions for Selected > Restart**.

### Configuring the Navigator Audit Server Data Expiration Period

**Required Role:** **Navigator Administrator** **Full Administrator**

You can configure the number of hours of audit events to keep in the Navigator Audit Server database as follows:

1. Do one of the following:

- Select **Clusters > Cloudera Management Service > Cloudera Management Service**.
- On the Status tab of the Home page, in **Cloudera Management Service** table, click the **Cloudera Management Service** link.

2. Click the **Configuration** tab.
3. Expand the **Navigator Audit Server Default Group** category.
4. Set the **Navigator Audit Server Data Expiration Period** property.
5. Click **Save Changes**.
6. Click the **Instances** tab.
7. Check the checkbox next to the **Navigator Audit Server** role.
8. Select **Actions for Selected > Restart**.

### Configuring the Audit Server to Mask Personally Identifiable Information

Required Role: **Navigator Administrator** **Full Administrator**

Personally identifiable information (PII) is information that can be used on its own or with other information to identify or locate a single person, or to identify an individual in context. The PII masking feature allows you to specify credit card number patterns (from major credit issuers) that are masked in audit events, in the properties of entities displayed in lineage diagrams, and in information retrieved from the Audit Server database and the Metadata Server persistent storage.

■ **Note:**

- Masking Social Security numbers is not supported in this release.
- Masking is not applied to audit events and lineage entities that existed before the mask was enabled.

1. Do one of the following:
  - Select **Clusters > Cloudera Management Service > Cloudera Management Service**.
  - On the Status tab of the Home page, in **Cloudera Management Service** table, click the **Cloudera Management Service** link.
2. Click the **Configuration** tab.
3. Expand the **Navigator Audit Server Default Group** category.
4. Click the **Advanced** category.
5. Configure the **PII Masking Regular Expression** property with a regular expression that matches the credit card number formats to be masked. The default expression is:

```
(4[0-9]{12}(?:[0-9]{3})?)|(5[1-5][0-9]{14})|(3[47][0-9]{13})|
(3(?:0[0-5]|[68][0-9])[0-9]{11})|(6(?:011|5[0-9]{2})[0-9]{12})|((?:2131|1800|35\d{3})\d{11})
```

which is constructed from the following subexpressions:

- Visa - (4[0-9]{12}(?:[0-9]{3})?)
- MasterCard - (5[1-5][0-9]{14})
- American Express - (3[47][0-9]{13})
- Diners Club - (3(?:0[0-5]|[68][0-9])[0-9]{11})
- Discover - (6(?:011|5[0-9]{2})[0-9]{12})
- JCB - ((?:2131|1800|35\d{3})\d{11})

If the property is left blank, PII information is not masked.

6. Click **Save Changes** to commit the changes.

## Cloudera Navigator Metadata Server

Describes how to add and configure the Navigator Metadata Server role.

- **Important: This feature is available only with a Cloudera Enterprise license.**

For other licenses, the following applies:

- Cloudera Express- The feature is not available.
- Cloudera Enterprise Data Hub Edition Trial - The feature is available until you end the trial or the trial license expires.

To obtain a license for Cloudera Enterprise, fill in this [form](#) or call 866-843-7207. After you install a Cloudera Enterprise license, the feature will be available.

## Adding the Navigator Metadata Server Role

**Required Role:** **Navigator Administrator** **Full Administrator**

Before adding the Navigator Metadata Server role, configure [the database](#) where policies, roles, and audit report metadata is stored.

1. Do one of the following:
  - Select **Clusters > Cloudera Management Service > Cloudera Management Service**.
  - On the Status tab of the Home page, in **Cloudera Management Service** table, click the **Cloudera Management Service** link.
2. Click the **Instances** tab.
3. Click the **Add Role Instances** button. The Customize Role Assignments page displays.
4. Assign the Navigator role to a host.
  - a. Customize the assignment of role instances to hosts. The wizard evaluates the hardware configurations of the hosts to determine the best hosts for each role. The wizard assigns all worker roles to the same set of hosts to which the HDFS DataNode role is assigned. These assignments are typically acceptable, but you can reassign them if necessary.

Click a field below a role to display a dialog containing a list of hosts. If you click a field containing multiple hosts, you can also select **All Hosts** to assign the role to all hosts or **Custom** to display the pageable hosts dialog.

The following shortcuts for specifying hostname patterns are supported:

- Range of hostnames (without the domain portion)

Range Definition	Matching Hosts
10.1.1.[1-4]	10.1.1.1, 10.1.1.2, 10.1.1.3, 10.1.1.4
host[1-3].company.com	host1.company.com, host2.company.com, host3.company.com
host[07-10].company.com	host07.company.com, host08.company.com, host09.company.com, host10.company.com

- IP addresses
- Rack name

Click the **View By Host** button for an overview of the role assignment by hostname ranges.

5. When you are satisfied with the assignments, click **Continue**. The Database Setup screen displays.
6. Configure database settings:
  - a. Choose the database type:
    - Leave the default setting of **Use Embedded Database** to have Cloudera Manager create and configure required databases. Make a note of the auto-generated passwords.

## Cluster Setup

### Database Setup

Configure and test database connections. If using custom databases, create the databases first according to the [Installing and Configuring an External Database](#) section of the [Installation Guide](#) .

☐ Use Custom Databases  
☒ Use Embedded Database

When using the embedded database, passwords are automatically generated. Please copy them down.

Database Name	Database Type	Database Name	Username	Password
<b>Hive</b> Database Host Name: tcdn53-1.ent.cloudera.com:7432	PostgreSQL	hive	hive	24FLyrj0zb
<b>Activity Monitor</b> Currently assigned to run on tcdn53-1.ent.cloudera.com. Database Host Name: tcdn53-1.ent.cloudera.com:7432	PostgreSQL	amon	amon	2VCic0tDJE
<b>Reports Manager</b> Currently assigned to run on tcdn53-1.ent.cloudera.com. Database Host Name: tcdn53-1.ent.cloudera.com:7432	PostgreSQL	rman	rman	Mn2l8tEoCH
<b>Navigator Audit Server</b> Currently assigned to run on tcdn53-1.ent.cloudera.com. Database Host Name: tcdn53-1.ent.cloudera.com:7432	PostgreSQL	nav	nav	P89RAR6e0o
<b>Navigator Metadata Server</b> Currently assigned to run on tcdn53-1.ent.cloudera.com. Database Host Name: tcdn53-1.ent.cloudera.com:7432	PostgreSQL	navms	navms	29O536GxZp

- Select **Use Custom Databases** to specify external databases.
  1. Enter the database host, database type, database name, username, and password for the database that you created when you set up the database.
- b. Click **Test Connection** to confirm that Cloudera Manager can communicate with the database using the information you have supplied. If the test succeeds in all cases, click **Continue**; otherwise check and correct the information you have provided for the database and then try the test again. (For some servers, if you are using the embedded database, you will see a message saying the database will be created at a later step in the installation process.) The Review Changes screen displays.

7. Click **Finish**.

## Starting, Stopping, and Restarting the Navigator Metadata Server

1. Do one of the following:
  - Select **Clusters > Cloudera Management Service > Cloudera Management Service**.
  - On the Status tab of the Home page, in **Cloudera Management Service** table, click the **Cloudera Management Service** link.
2. Click the **Instances** tab.
3. Do one of the following depending on your role:
  - **Required Role:** **Full Administrator**
    1. Check the checkbox next to the **Navigator Metadata Server** role.
    2. Select **Actions for Selected > Action**. Click **Action** to confirm the action, where *Action* is Start, Stop, or Restart.
  - **Required Role:** **Navigator Administrator** **Full Administrator**

1. Click the **Navigator Metadata Server** role link.
2. Select **Actions** > **Action this Navigator Metadata Server**. Click **Action this Navigator Metadata Server**, where *Action* is Start, Stop, or Restart, to confirm the action.

### Configuring the Navigator Metadata Server Storage Directory

**Required Role:** **Navigator Administrator** **Full Administrator**

Describes how to configure where the Navigator Metadata Server stores extracted data. The default is `/var/lib/cloudera-scm-navigator`.

1. Do one of the following:
  - Select **Clusters** > **Cloudera Management Service** > **Cloudera Management Service**.
  - On the Status tab of the Home page, in **Cloudera Management Service** table, click the **Cloudera Management Service** link.
2. Click the **Configuration** tab.
3. Click the **Navigator Metadata Server Default Group**.
4. Specify the directory in the **Navigator Metadata Server Storage Dir** property.
5. Click **Save Changes**.
6. Click the **Instances** tab.
7. Check the checkbox next to the **Navigator Metadata Server** role.
8. Select **Actions for Selected** > **Restart**.

### Configuring the Navigator Metadata Server Port

**Required Role:** **Navigator Administrator** **Full Administrator**

Describes how to configure the port on which the Navigator UI is accessed. The default is 7187.

1. Do one of the following:
  - Select **Clusters** > **Cloudera Management Service** > **Cloudera Management Service**.
  - On the Status tab of the Home page, in **Cloudera Management Service** table, click the **Cloudera Management Service** link.
2. Click the **Configuration** tab.
3. Select **Navigator Metadata Server Default Group** > **Ports and Addresses**.
4. Specify the port in the **Navigator Metadata Server Port** property.
5. Click **Save Changes**.
6. Click the **Instances** tab.
7. Check the checkbox next to the **Navigator Metadata Server** role.
8. Select **Actions for Selected** > **Restart**.

### Navigator Metadata Server Sizing and Performance Recommendations

**Required Role:** **Navigator Administrator** **Full Administrator**

Two activities determine Navigator Metadata Server resource requirements:

- Extracting metadata from the cluster and creating relationships
- Querying

The Navigator Metadata Server uses Solr to store, index, and query metadata. Indexing happens during extraction. Querying is fast and efficient because the data is indexed.

Memory and CPU requirements are based on amount of data that is stored and indexed. With 6 GB of RAM and 8-10 cores Solr can process 6 million entities in 25-30 minutes or 80 million entities in 8 to 9 hours. Any less RAM than 6GB and will result in excessive garbage collection and possibly out-of-memory exceptions. For large clusters, Cloudera advises at least 8 GB of RAM and 8 cores. The Solr instance runs in process with Navigator, so the Java heap for the Navigator Metadata Server should be set according to the size of cluster.

By default, during the Cloudera Manager Installation wizard the Navigator Audit Server and Navigator Metadata Server are assigned to the same host as the Cloudera Management Service monitoring roles. This configuration works for a small cluster, but should be updated before the cluster grows. You can either change the configuration at installation time or move the Navigator Metadata Server if necessary.

### Moving a Navigator Metadata Server Role

**Required Role:** Navigator Administrator Full Administrator

1. Stop the Navigator Metadata Server role, delete it from existing host, and add it to a new host.
2. If the Solr data path is not on NFS/SAN, move the data to the same path on the new host.
3. Start the Navigator Metadata Server role.

### Enabling Hive Metadata Extraction in a Secure Cluster

**Required Role:** Navigator Administrator Full Administrator

The Navigator Metadata Server uses the hue user to connect to the Hive Metastore. The hue user is able to connect to the Hive Metastore by default. However, if the Hive service **Hive Metastore Access Control and Proxy User Groups Override** property and/or the HDFS service **Hive Proxy User Groups** property have been changed from their default values to settings that prevent the hue user from connecting to the Hive Metastore, Navigator Metadata Server will be unable to extract metadata from Hive. If this is the case, modify the Hive service **Hive Metastore Access Control and Proxy User Groups Override** property and/or the HDFS service **Hive Proxy User Groups** property so that the hue user can connect as follows:

1. Go to the Hive or HDFS service.
2. Click the **Configuration** tab.
3. Expand the **Service-Wide > Proxy** category.
4. In the Hive service **Hive Metastore Access Control and Proxy User Groups Override** field or the HDFS service **Hive Proxy User Groups** field, click the Value column, and click **+** to add a new row.
5. Type `hue`.
6. Click **Save Changes** to commit the changes.
7. Restart the service.

### Configuring the Metadata Server to Mask Personally Identifiable Information

**Required Role:** Navigator Administrator Full Administrator

Personally identifiable information (PII) is information that can be used on its own or with other information to identify or locate a single person, or to identify an individual in context. The PII masking feature allows you to specify credit card number patterns (from major credit issuers) that are masked in audit events, in the properties of entities displayed in lineage diagrams, and in information retrieved from the Audit Server database and the Metadata Server persistent storage.

■ **Note:**

- Masking Social Security numbers is not supported in this release.
- Masking is not applied to audit events and lineage entities that existed before the mask was enabled.



1. Do one of the following:
  - Select **Clusters > Cloudera Management Service > Cloudera Management Service**.
  - On the Status tab of the Home page, in **Cloudera Management Service** table, click the **Cloudera Management Service** link.
2. Click the **Configuration** tab.
3. Expand the **Navigator Metadata Server Default Group** category.
4. Click the **Advanced** category.
5. Configure the **PII Masking Regular Expression** property with a regular expression that matches the credit card number formats to be masked. The default expression is:

```
(4[0-9]{12}(?:[0-9]{3})?)|(5[1-5][0-9]{14})|(3[47][0-9]{13})|
(3(?:0[0-5]|[68][0-9])[0-9]{11})|(6(?:011|5[0-9]{2})[0-9]{12})|((?:2131|1800|35\\d{3})\\d{11})
```

which is constructed from the following subexpressions:

- Visa - (4[0-9]{12}(?:[0-9]{3})?)
- MasterCard - (5[1-5][0-9]{14})
- American Express - (3[47][0-9]{13})
- Diners Club - (3(?:0[0-5]|[68][0-9])[0-9]{11})
- Discover - (6(?:011|5[0-9]{2})[0-9]{12})
- JCB - ((?:2131|1800|35\\d{3})\\d{11})

If the property is left blank, PII information is not masked.

6. Click **Save Changes** to commit the changes.

### Configuring a JMS Server for Policy Messages

1. Do one of the following:
  - Select **Clusters > Cloudera Management Service > Cloudera Management Service**.
  - On the Status tab of the Home page, in **Cloudera Management Service** table, click the **Cloudera Management Service** link.
2. Click the **Configuration** tab.
3. Expand the **Navigator Metadata Server Default Group** category.
4. Expand the **Policies** category.
5. Set the following properties:

Property	Description
JMS URL	The URL of the JMS server to which notifications of changes to entities affected by policies are sent. Default: tcp://localhost:61616.
JMS User	The JMS user to which notifications of changes to entities affected by policies are sent. Default: Navigator.
JMS Password	The password of the JMS user to which notifications of changes to entities affected by policies are sent. Default: admin.
JMS Queue	The JMS queue to which notifications of changes to entities affected by policies are sent. Default: admin.

## Cloudera Navigator Administration

6. Click **Save Changes** to commit the changes.
7. Restart the Metadata Server role.

### Enabling and Disabling Policy Expressions

Required Role: **Navigator Administrator** **Full Administrator**

1. Do one of the following:
  - Select **Clusters > Cloudera Management Service > Cloudera Management Service**.
  - On the Status tab of the Home page, in **Cloudera Management Service** table, click the **Cloudera Management Service** link.
2. Click the **Configuration** tab.
3. Expand the **Navigator Metadata Server Default Group** category.
4. Expand the **Policies** category.
5. Check or uncheck the **Enable Expression Input** checkbox.
6. Click **Save Changes** to commit the changes.
7. Restart the Metadata Server role.


## Displaying the Cloudera Navigator Version

To display the version and build number for Cloudera Navigator:

1. [Start and log into the Navigator UI](#).
2. Select  > **About**.

## Displaying Cloudera Navigator Documentation

To display Cloudera Navigator documentation:

1. [Start and log into the Navigator UI](#).
2. Select  > **Help**. The Cloudera Navigator online documentation displays in a new window.