Hands-On Exercise: Configuring a Hadoop Cluster

In this exercise, you will modify a service configuration, activate additional parcels, and add additional services to your cluster. You will then create a Cloudera Manager host template and apply it to an existing host in the cluster.

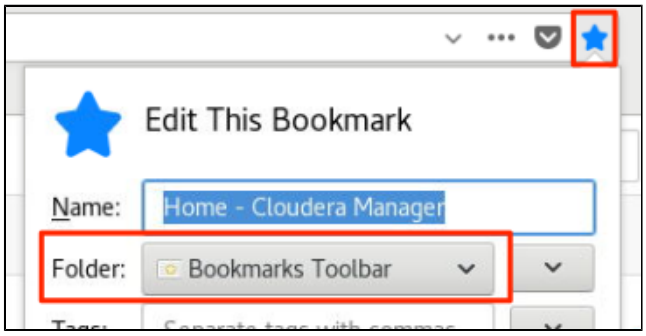
Modify a service configuration

In this step, you will practice changing configuration settings by changing the HDFS replication factor. (The replication factor determines how many copies of each file in HDFS are kept on the cluster.)

1. Locate and change the HDFS replication factor setting.

   a. In the remote desktop browser, go to the Cloudera Manager admin console home page (`cmhost:7180`).

      Tip: You might wish to bookmark the Cloudera Manager home page, because you will visit it frequently throughout the exercises.



   b. Go the HDFS service page by clicking the **HDFS** service in the **Cluster 1** list of services.

   c. Select the **Configuration** tab.

   d. From the **Category** set of filters on the left, select the **Replication** filter.

   e. Change the replication Factor to **2**.

   f. Click **Save Changes**.

---

2. Restart stale services and redeploy client configurations:

a. Return to the CM home page by clicking the Cloudera Manager logo in the upper left corner.

b. Note that the HDFS service (and several services that depend on HDFS) show two new status icons. One is the "Stale Configuration: Restart needed" icon. The other is the "Stale Configuration: Client configuration redeployment needed" icon.

Click on either one of the two icons. This opens the **Stale Configurations** page.

c. Review the changes that Cloudera Manager will push out to the cluster hosts and then click **Restart Stale Services**.

d. In the **Review Changes** screen, keep **Re-deploy client configuration** checked and click **Restart Now**.

e. The **Restart Awaiting Staleness Computation Command** page is displayed. Wait for the commands to complete, which should take less than five minutes. Then click **Finish**.

> **Resolving Stale Configurations**
> In future exercises, you might be instructed to change configuration settings without being explicitly reminded to redeploy the changes. Whenever you see stale configuration icons, you can resolve them as you did above.

Allocate more memory to YARN Resource settings

The default YARN resources setting is insufficient to run all configured services. In this section, you will increase YARN's resources to accommodate the installed services for this cluster.

3. In Cloudera Manager, select the YARN service.

---

4. Select the **Configuration** tab.

---

5. In the search bar, enter **resource_memory**.

---

6. Under **NodeManager Default Group**, enter **3 GiB**.

7. Review the changes in Cloudera Manager, and click **Restart Stale Services**.

Set an unexposed property using a configuration snippet

Most configuration properties can be set directly in Cloudera Manager. However, some less common properties are not exposed, and must be set using a "safety valve" snippet, which modifies a configuration file directly.

One such property is `dfs.datanode.scan.period.hours`, which determines how often HDFS scans for corrupt files.

8. In Cloudera Manager, return to the HDFS **Configuration** tab.

9. Find the **HDFS Service Advanced Configuration Snippet (Safety Valve) for hdfs-site.xml** property.

   Tip: You can find this either by selecting **Category > Advanced** in the **Filters** panel on the left, or by entering `hdfs-site` in the search box.

10. Notice that the advanced configuration snippet currently contains no settings. Click the plus sign icon next to the configuration name to add a setting.

11. In the **Name** field, enter `dfs.datanode.scan.period.hours`.

12. In the **Value** field, enter `240`. This overrides the default scan frequency (every three weeks) to be every 10 days.

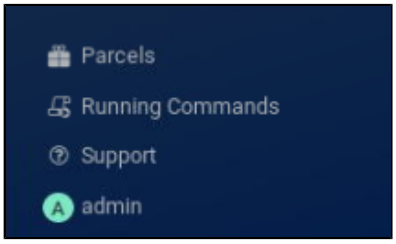
13. *Optional*: Enter a description of the change, such as `scan every 10 days`.

14. Click **View as XML**. This displays the actual command that will be added to the `hdfs-site.xml`.

15. As you did in the previous section, save your changes, then redeploy the configuration files and restart the affected services.

Distribute and activate parcels

16. Go to the **Parcels** page in Cloudera Manager by clicking on the parcel icon on the bottom left.



---

17. Under **Location** in the left filter panel, ensure that **Cluster 1** (the name of your cluster) is highlighted.

---

18. Click **Distribute** for the CFM parcel *only*. Do *not* touch the 7.1.6 CDH parcel.

---

19. When CM has distributed the CFM parcel, click **Activate**. If you are prompted for whether you are sure want to activate the parcel, click **OK**. Once the parcel is activated, the button will display **Deactivate**. Leave it in that stage.

---

20. If any of the cluster services show stale configuration icons, resolve the stale configuration as you did earlier.

Modify YARN configuration

The following steps will help keep a healthy status for Hive on Tez in your exercise environment.

21. Log into CM with the `admin` credentials.

---

22. Go to **Configuration**.

---

23. Set the max *per-container* RAM and CPU allocations:

    a. Set `yarn.scheduler.maximum-allocation-mb` to **4 GiB**.

    b. Set `yarn.scheduler.maximum-allocation-vcores` to **1**.

---

24. Set the max total RAM and CPU allocations:

a. Set `yarn.nodemanager.resource.memory-mb` to **8 GiB**.

b. Set `yarn.nodemanager.resource.cpu-vcores` to **8**.

---

25. Save the changes and restart the services.

Add NiFi to the cluster

26. From the Cloudera Manager home page, choose **Add Service** from your cluster drop-down menu.

---

27. Choose **NiFi** and click **Continue**.

---

28. On the next screen, assign the **NiFi Node** role to **master-2**. Click **Continue**.

---

29. Continue through the rest of the wizard, keeping all default settings, until NiFi is installed. Click **Finish**.

---

30. If any of the cluster services show stale configuration icons, resolve the stale configuration as you did earlier in this exercise.

---

31. NiFi will appear as it is running, and two health issues will appear on the cluster status page associated with NiFi.

This problem is a known issue with this version of Nifi running on CDP 7.1.5. Upon research you would find a workaround as explained in the next steps.

---

32. Locate the advanced configuration snippet again for `staging/state-management.xml` and add two entries:

a. Click **NiFi** from the services list, and then click the **Configuration** tab.

b. In the search box, enter **Node Advanced Configuration Snippet (Safety Valve) for staging/state-management.xml**.

c. Notice that the advanced configuration snippet currently contains no settings. Click on the plus sign to add a new setting.

d. Add a new setting with the following values:

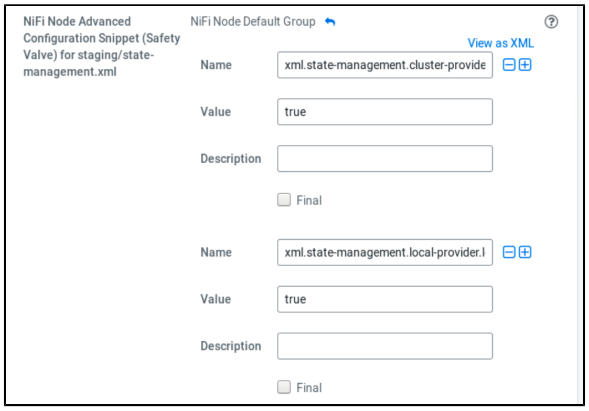
Name: `xml.state-management.cluster-provider.zk-provider.enabled`

Value: `true`

e. Add another new setting with the following values:

Name: `xml.state-management.local-provider.local-provider.enabled`

Value: `true`



f. Click **Save Changes**. Restart the stale services and redeploy client configurations.

Add Sqoop 1 to the cluster

Sqoop is a client install only. We will add the service, then add a Sqoop gateway in the next exercise when we create a gateway template.

33. From the Cloudera Manager home page, choose **Add Service** from your cluster drop-down menu.

---

34. Choose **Sqoop** and click **Continue**.

---

35. Do not deploy the **Gateway** role to any host at this time.

---

36. Continue with default settings through the rest of the wizard.

Create a host template for gateway hosts

The `cmhost` machine is currently a utility node. In this step, you will configure it with gateway roles so that it also plays the role of a gateway (or edge) node. This setup is consistent with Cloudera's recommendation for a cluster with fewer than 20 hosts.

37. Go to **Hosts** > **Host Templates**.

---

38. Click **Create**.

---

39. Enter template name `Gateway`.

---

40. Expand the HDFS and check the **Gateway**. Leave the gateway group set to **Gateway Default Group**.

Continue this process to add the Gateway roles for each of the following services as well:

- HDFS
- Hive
- Hive-on-Tez
- Sqoop
- YARN

---

41. Click **Create** to save the template.

---

42. Verify the settings of the **Gateway** template: five **Gateway Default Group** roles listed in the **Groups** column.

Apply the Gateway host template to `cmhost`

43. Select **Hosts** > **All Hosts**.

---

44. Check the box next to `cmhost`.

---

45. From the **Actions for Selected** menu, choose **Apply Host Template**.

---

46. Choose the **Gateway** host template you just created.

---

47. Place a checkmark next to **Deploy client configurations...**

---

48. Click **Confirm**.

---

49. The gateway role instances identified by the template you created will be deployed to `cmhost`. This will take a few minutes to complete.

When it completes, click **Close**.

---

50. Return to the Cloudera Manager home page.

---

51. If any of the cluster services show stale configuration icons, resolve the stale configuration.

---

52. Return to the **Hosts** > **All Hosts** page. Then click on blue arrow icon in the **Roles** column for `cmhost`.

Notice that `cmhost` now hosts the gateway roles that are part of the **Gateway** template, in addition to the other roles assigned to it earlier.

Tip: The gateway roles do not have running indicators (green checkmark icons) or stopped indicators because the gateway roles do not include daemons. They are client libraries that allow users to interact with the services to which they connect.

This is the end of the exercise.