



Cloudera Administrator Training for CDP PVC Base: Hands-On Exercises

Table of Contents

General Notes	1
Hands-On Exercise: Verify the Exercise Network	5
Hands-On Exercise: Installing Cloudera Manager Server	8
Hands-On Exercise: Cluster Installation	16
Hands-On Exercise: Configuring a Hadoop Cluster	26
Hands-On Exercise: Working with HDFS	34
Hands-On Exercise: Storing Data in Amazon S3	47
Hands-On Exercise: Importing Data Using Sqoop	48
Hands-On Exercise: NiFi Verification	52
Hands-On Exercise: Working with Kafka	56
Hands-On Exercise: Install Impala and Hue	60
Hands-On Exercise: Using Hue, Hive and Impala	62
Hands-On Exercise: Running YARN Applications	71
Hands-On Exercise: Running Spark Applications	76
Hands-On Exercise: Using The Capacity Scheduler	79
Hands-On Exercise: Configuring HDFS for High Availability	83
Hands-On Exercise: Creating and Using a Snapshot	87

Hands-On Exercise: Upgrade the Cluster	93
Hands-On Exercise: Breaking the Cluster	96
Hands-On Exercise: Confirm Cluster Healing and Configuring Email Alerts	98
Hands-On Exercise: Troubleshooting a Cluster	102

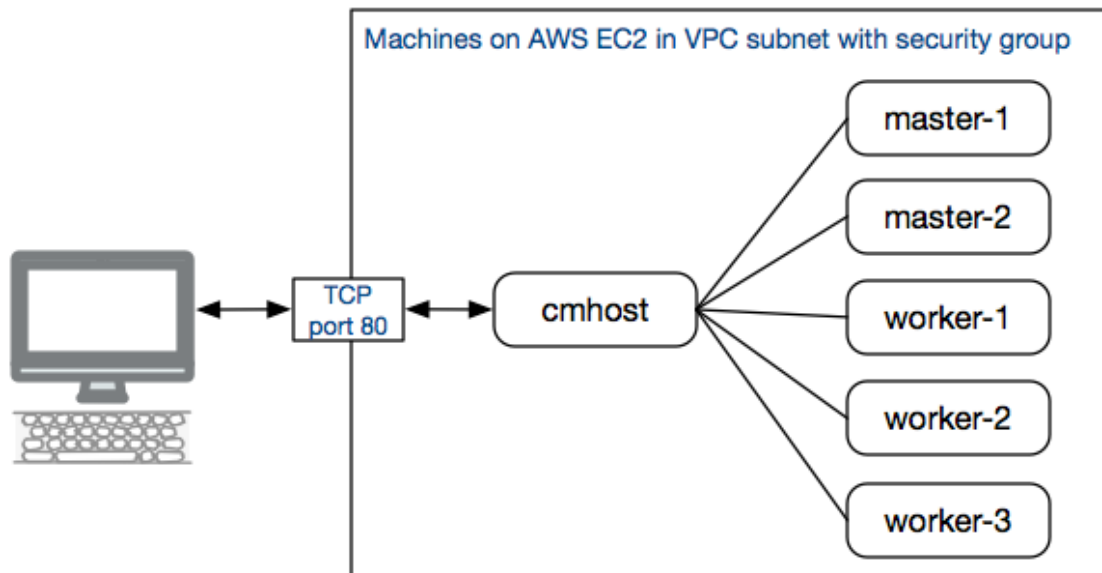
General Notes

Important: Read this section before starting the exercises.

Using Your Exercise Environment

Exercise Environment Overview

In this course, you will install Cloudera Manager and CDP on six virtual machines (VMs) running in the cloud. These are referred to as the “cluster hosts.”



Logging in to the Exercise Environment

All exercise steps are performed on the cmhost desktop, using the Firefox browser and terminal window(s).

Cluster host login credentials (all hosts):

- Username: training
- Password: training

Superuser access: You can use `sudo` without entering a password. The training user has unlimited, passwordless `sudo` privileges.

Logging in to other hosts: You can use `ssh` to connect from cmhost to the other cluster hosts without entering a password.

Note: Passwordless access is enabled for convenience in the course environment, but should not be enabled in a production environment.

Points to Note While Doing the Exercises

Directories

The main directory for the exercises is `~/training_materials/admin`.

The `admin` directory contains these subdirectories:

- `data`—data files used in the exercises
- `java`—Java applications you will run during the exercises
- `scripts`—scripts you will run during the exercises
- `solutions`—text files containing the commands you will need to enter during the exercises
 - Use the solution files to make copying and pasting text easier.

Fewer Step-by-Step Instructions as You Work Through These Exercises

As the exercises progress, and you gain more familiarity with the tools and environment, we will provide fewer step-by-step instructions. As in the real world, we merely give you a requirement, and it is up to you to solve the problem. You should feel free to refer to the hints or solutions provided, ask your instructor for assistance, or consult with your fellow students.

Notational Convention

In some command-line steps in the exercises, you will see lines like this:

```
$ hdfs dfs -put shakespeare \  
/user/training/shakespeare
```

The backslash at the end of the first line signifies that the command is not complete and continues on the next line. You can enter the code exactly as shown (on multiple lines), or you can enter it on a single line and remove the backslashes.

Copying and Pasting Command Line Text from Exercise Manual

Sometimes you might need to copy commands from the exercises and paste them into your terminal window. There are two options to do this:

- Copy from the provided solution text files in the `training_materials/admin/solutions` directory. There is one text file for each exercise.

- Copy directly from the instructions.

However, note that copying multi-line commands from PDF to text can sometimes introduce problems. If you copy a line that ends a backslash (\), sometimes it will append the following line immediately after the backslash. If this happens, you need to manually remove the backslash.

Reset Script—Resetting Your Cluster

For most exercises, if you did not successfully complete the prerequisite exercise(s), **you can use the `reset_cluster.sh` script to set the state of your cluster as necessary to perform an exercise.**

```
$ ~/training_materials/admin/scripts/reset_cluster.sh
```

Note the following points regarding the reset script:

- *Networking between hosts in your environment **must** be configured correctly before running the reset script. Be sure you have completed [Hands-On Exercise: Configuring a Hadoop Cluster](#) at least once before running the reset script. It doesn't need to be ran each time the reset script is ran.*
- Some prerequisite exercises *cannot* be bypassed using a script. If you are unable to complete those exercises, speak to your instructor.
- The reset script is destructive: work you have already done on the cluster might be overwritten when the script runs.
- You can reset your environment to the original state, before Cloudera Manager is installed, by selecting **Full cluster reset**.
- Depending on which environment state you are resetting to, the script may take 30 minutes or more.
- The options you will be given include:
 - Full cluster reset - CM installed, no cluster
 - Cluster setup - CM and Basic cluster installed (thru ch 3)
 - High-availability setup - (thru ch 12)
 - Final setup - end of course (thru ch 15)

- After the cluster restarts, you might see some services with an orange dot indicating a concerning state due to swapping on cmhost. The warnings should resolve in about 15 minutes. You can continue with the exercises in the meantime.

Reset Data Script—Resetting Your Datasets

For most exercises, if you did not successfully complete the prerequisite exercise(s), **you can use the `reset_cluster.sh` script to set the state of your cluster as necessary to perform an exercise.**

```
$ ~/training_materials/admin/scripts/  
inter_active_dataset_reset.sh
```

Note the following points regarding the reset script:

- *Networking between hosts in your environment **must** be configured correctly before running the cluster reset script. If the cluster reset script needs to be ran, be sure you have completed [Hands-On Exercise: Configuring a Hadoop Cluster](#) before running the reset dataset script.*
- The options you will be given include:
 - Full tables and data clean up
 - Manage datasets on HDFS
 - Import data to HDFS with Sqoop
 - Create and manage external Hive tables

Hands-On Exercise: Verify the Exercise Network

In this exercise, you will verify your exercise environment's network in preparation for installing Cloudera Manager and deploying a Hadoop cluster.

IMPORTANT: Be sure to read through [General Notes](#) before starting this exercise.

1. Open a terminal prompt on the cmhost remote desktop.
 - From the desktop **Applications** menu, choose **System Tools** > **Mate Terminal**.
 - You should see a terminal window with a `training@cmhost:~$` prompt.
-

2. Verify that you can communicate with all the hosts in your cluster from cmhost by using the hostnames.

In the cmhost terminal, enter:

```
training@cmhost:~$ ping -c 3 cmhost
training@cmhost:~$ ping -c 3 master-1
training@cmhost:~$ ping -c 3 master-2
training@cmhost:~$ ping -c 3 worker-1
training@cmhost:~$ ping -c 3 worker-2
training@cmhost:~$ ping -c 3 worker-3
```

If ping returns an error that the host is unreachable, this means the network for the VM is down. Use the following steps to restart the VM network.

- a. Return to your Exercise Environment portal (showing the six machines for this course). Click the machine whose network is down.
- b. Login as user `training` with password `training`.
- c. Open a terminal window (double click the computer icon on the desktop, or click the computer icon in the menu bar) and run this command to restart the network:

```
$ sudo systemctl restart network
```

3. Confirm that the `training` user can use `ssh` to connect to the other virtual machines.

Note: Your environment is configured so that you can log in, submit commands, and copy files from `cmhost` to the other hosts in the environment without a password.

Passwordless SSH and `scp` **are not required to deploy a CDP cluster**. Passwordless connection is configured in the classroom environment as a convenience—do not do this in a production environment.

- In the same terminal window, connect to the `master-1` machine:

```
training@cmhost:~$ ssh master-1
```

- Confirm that the terminal window prompt has changed to `training@master-1:~$`, indicating that you are now connected to `master-1`.

Note: You might get a message starting with `Warning: Permanently added...` You can disregard this message.

- Disconnect from the `master-1` machine.

```
training@master-1:~$ exit
```

- Confirm that the terminal window prompt has changed back to `training@cmhost:~$`.
- Use the same approach to connect to each of the other virtual machines. One at a time, connect to each of the other machines that will be part of the cluster, then exit each one after confirming that the prompt displays as show below.

```
training@cmhost$ ssh master-2
training@master-2:~$ exit
training@cmhost:~$ ssh worker-1
training@worker-1:~$ exit
training@cmhost:~$ ssh worker-2
training@worker-2:~$ exit
training@cmhost:~$ ssh worker-3
training@worker-3:~$ exit
```

Note: To simplify the instructions, the rest of the exercises will use a dollar sign (\$) instead of the full prompt (such as


```
training@cmhost:~$) to indicate the cmhost terminal  
prompt.
```

This is the end of the exercise.

Hands-On Exercise: Installing Cloudera Manager Server

For this installation, you will install Cloudera Manager Server on the `cmhost` machine. Before installing Cloudera Manager, you will configure an external database (MySQL) to be used by Cloudera Manager and some of the services, which you will install in the next exercise.

Verify Environment Configuration

IMPORTANT: Complete all steps in this exercise on `cmhost`

In this section of the exercise, you will verify some important settings prior to installing software. Although in these steps you will only verify settings on the `cmhost` machine, most of the settings below should be in place on all the machines that will be part of the cluster.

1. Test the repo URL that contains files to install Cloudera Manager. From `cmhost`, start a web browser and enter this URL: <http://cmhost:8060/cm7.3.1/>
2. Verify the Oracle JDK is installed and that `JAVA_HOME` is defined and referenced in the system `PATH`.

In a terminal:

```
$ java -version
```

The message returned in the terminal should show the java version is 1.8.0_232.

```
$ echo $JAVA_HOME
```

The result should show that `/usr/java/default` is the `JAVA_HOME` location.

```
$ env | grep PATH
```

The `PATH` value returned includes a reference to `/usr/java/default/bin`.

3. Verify Python is installed. It is a requirement for Hue, which you will install later in the course.

```
$ rpm -q python-2.7*
```

The command should return the package name of the installed version of python 2.7.

4. Verify MySQL Server is installed and running on cmhost.

```
$ systemctl status mysqld
```

The results of the command should show that the `mysqld.service` service is “active” (running).

Note: Note that in a true production deployment you would also move the old InnoDB log files to a backup location and update the `/etc/my.cnf` MySQL configurations to conform with requirements as documented [here](#).

5. Confirm no blocking is being done by Security-Enhanced Linux (SELinux)

```
$ sestatus
```

Note: Although it is not a requirement to set SELinux to disabled or permissive, it is important that SELinux not block during installation.

6. Verify IPv6 is disabled. This is a CDP requirement.

```
$ ip addr show
```

Notice that there is an IPv4 (`inet`) address for the `eth0` network interface, however there is no `inet6` address.

7. Check firewall settings.

```
$ systemctl list-unit-files | grep firewalld
```

The results of the command should show that the `firewalld.service` service is “disabled”.

Note: Although it is not a requirement to disable the firewall, there are many ports that must not be blocked during and after installation.

8. Check Transparent Hugepage compaction is disabled.

```
$ cat /proc/meminfo | grep -i hugepages_total
```

A response of HugePages_Total: 0 indicates the hugepage feature is turned off.

Note: The OS feature called transparent hugepages interacts poorly with Hadoop workloads and can seriously degrade performance. Cloudera recommends it be turned off.

9. Check the vm.swappiness Linux kernel setting.

```
$ cat /proc/sys/vm/swappiness
```

Note: Cloudera recommends that you set vm.swappiness to a value between 1 and 10, in order to reduce lengthy garbage collection pauses which can affect stability and performance.

10. Verify the service that ensures time consistency across machines is running.

```
$ ntpstat
```

The results of the command should show that the time is synchronized with the ntp server.

11. Verify the MySQL JDBC Connector is installed. Sqoop (a part of CDP that you will install in this course) does not ship with a JDBC connector, but does require one.

```
$ ls -l /usr/share/java/mysql*
```

You should see that a symlink has been defined at /usr/share/java/mysql-connector-java.jar that points to a specific version of a MySQL connector JAR file that exists in the same directory.

Configure the External Cloudera Manager Database

As is typical in a production cluster, Cloudera Manager uses an external database system instead of the embedded PostgreSQL system.

12. Optional: Review the script you will run to create the required databases.

```
$ cat ~/training_materials/admin/scripts/mysql-setup.sql
```

This script creates databases and users for the Cloudera Manager database and databases required by other services on the cluster.

13. Run the script to create the databases.

```
$ mysql -u root -p < ~/training_materials/admin/scripts\
/mysql-setup.sql
```

When prompted for the password enter **training**

14. Confirm the databases were created.

```
$ mysql -u root -p -e "SHOW DATABASES"
```

When prompted for the password enter **training**

Confirm that there are 13 databases, include these that were created by the script: amon, cmserver, hue, metastore, oozie, registry, rman, and streamsmgmgr. (The list will also include several previously created databases.)

15. Make your MySQL installation secure.

```
$ sudo /usr/bin/mysql_secure_installation
[...]
Enter current password for root (enter for none): training
OK, successfully used password, moving on..
[...]
Change the root password? [Y/n] N
Remove anonymous users? [Y/n] Y
[...]
Disallow root login remotely? [Y/n] Y
[...]
Remove test database and access to it [Y/n] Y
[...]
Reload privilege tables now? [Y/n] Y
All done!
[...]
Thanks for using MySQL!
Cleaning up . . .
```

16. Verify the Cloudera Manager local software repository.

Your instances contain a local yum repository of Cloudera Manager software to save download time in this course.

CentOS (and Red Hat) store software repository references in `/etc/yum.repos.d`.

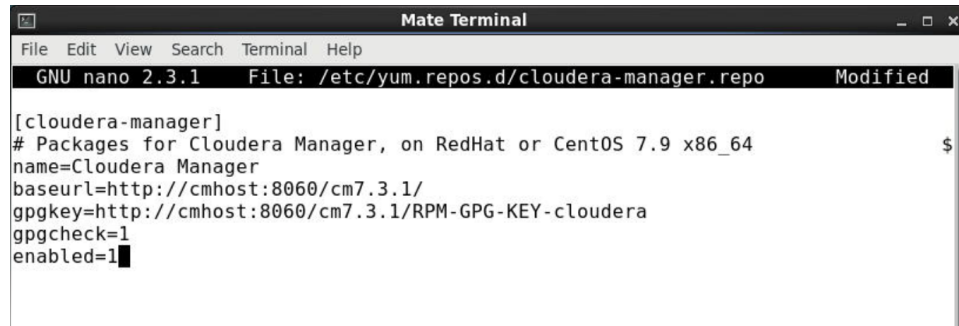
- Issue the commands below to edit the yum repository settings.

```
$ sudo cp ~/config/cloudera-manager.repo /etc/yum.repos.d/
$ sudo nano /etc/yum.repos.d/cloudera-manager.repo
```

- Update the file with the following settings:

```
[cloudera-manager]
baseurl = http://cmhost:8060/cm7.3.1/
enabled = 1
gpgcheck = 1
gpgkey = http://cmhost:8060/cm7.3.1/RPM-GPG-KEY-cloudera
name = Cloudera Manager
```

Notice the base URL in the output of the last command. You will type this URL into Cloudera Manager later when you install Cloudera Manager agents.



```

Mate Terminal
File Edit View Search Terminal Help
GNU nano 2.3.1 File: /etc/yum.repos.d/cloudera-manager.repo Modified

[cloudera-manager]
# Packages for Cloudera Manager, on RedHat or CentOS 7.9 x86_64
name=Cloudera Manager
baseurl=http://cmhost:8060/cm7.3.1/
gpgkey=http://cmhost:8060/cm7.3.1/RPM-GPG-KEY-cloudera
gpgcheck=1
enabled=1
  
```

Then save the file using `Ctrl + O`, and hit enter to accept the location. Exit with `Ctrl + X`.

Install Cloudera Manager Server

17. Install Cloudera Manager Server.

```

$ cd ~/software/cm7.3.1/RPMS/x86_64
$ sudo yum localinstall -y cloudera-manager-daemons-7* \
  cloudera-manager-server-7*
  
```

Note: The `-y` option provides an answer of yes in response to an expected confirmation prompt.

Note: This command will take some time due to the size of the new CDP RPM packages.

18. Run the script to prepare the Cloudera Manager database.

```

$ sudo /opt/cloudera/cm/schema/scm_prepare_database.sh \
  mysql cmserver cmserveruser password
  
```

After running the command above you should see the message, “All done, your SCM database is configured correctly!”

19. Start the Cloudera Manager Server.

```

$ sudo systemctl start cloudera-scm-server
  
```

20. Confirm that Cloudera Manager Server process started successfully.

```
$ ps -ef | grep cloudera-scm-server
```

The results of the `ps` command above show that Cloudera Manager Server is using the JDBC MySQL connector to connect to MySQL. It also shows logging configuration and other details.

Note: You should see multiple entries for the `cloudera-scm-server` process. If there is only one entry for the `cloudera-scm-server` process that displays, it means that the service did not start correctly. This can usually be corrected by rebooting `cmhost`. Use `sudo reboot` and enter password `training` when prompted. After rebooting, you will need to reconnect your browser session to `cmhost`.

21. Review the Cloudera Manager Server log file to see what took place.

The path to the log file is `/var/log/cloudera-scm-server/cloudera-scm-server.log`. Note that you must use `sudo` to access Cloudera Manager logs because of restricted permissions on the Cloudera Manager log file directories.

```
$ sudo tail -f /var/log/cloudera-scm-server/cloudera-scm-server.log
```

Tip: Bash tab-completion won't work for the filename in the above command, because the permissions on the directory do not allow the training user to view its contents until the `sudo` command actually runs.

Tip: Press the Enter key to scroll through the log file contents. Press the `Ctrl + c` to quit the `less` command file viewer.

This is the end of the exercise.

Hands-On Exercise: Cluster Installation

In this exercise, you will log in to the Cloudera Manager Admin Console to install, deploy and verify your cluster.

During this exercise, you will identify the hosts that Cloudera Manager will manage.

You will then be prompted to choose which CDP services you want to add in the cluster and to which machines you would like to add each role of each service.

At the end of this exercise, the Cloudera Manager and CDP services and roles will be deployed across your cluster as show below. (The services and roles added in this exercise are in blue).

	master-1	master-2	worker-1	worker-2	worker-3	cmhost
HDFS NameNode	✓					
HDFS Secondary NameNode		✓				
HttpFS	✓					
HDFS Balancer	✓					
HDFS DataNode			✓	✓		✓
Hive Metastore	✓					
HiveServer 2 on Tez	✓					
Oozie Server		✓				
Spark History Server		✓				
Spark Gateway	✓	✓				
YARN Resource Manager	✓					
YARN JobHistory Server	✓					
YARN NodeManager		✓	✓	✓		✓
Zookeeper Server	✓	✓	✓			
Cloudera Manager Server						✓
Cloudera Manager Server Database						✓
Cloudera Management Services						✓
Cloudera Manager Agent	✓	✓	✓	✓		✓

In subsequent exercises, you will add more services to your cluster.

After completing the installation steps, you will review a Cloudera Manager Agent log file and review processes running on a machine in the cluster.

All steps in this exercise that use a terminal window should be run on cmhost.

Log in to Cloudera Manager Admin UI

1. If Firefox is not yet running, launch it from the cmhost desktop's **Applications, Internet, Firefox**.

Note: Firefox may take a minute to launch the first time it is started.

2. In Firefox, load the Cloudera Manager Admin Console by entering `http://cmhost:7180`.

Note: If an "Unable to Connect" message appears, the Cloudera Manager server has not yet fully started. Wait several moments, and then attempt to connect again.

3. Log in with the username `admin` and password `admin`.

The **Welcome to Cloudera Manager** page.

4. On the "Welcome to Cloudera Manager 7.3.1" page, select **Try Cloudera Data Platform for 60 days**. Add a checkmark in the box next to **Yes, I accept the Cloudera Standard License Terms and Conditions** and click **Continue**.
-

5. The **Add Cluster - Installation** page. Click **Continue**.
-

6. The **Cluster Basics** page will appear. Enter the cluster name or accept the default Cluster 1. Click **Continue**. A Regular Cluster will be created.
-

Install Cloudera Manager agents and create cluster

7. The **Specify Hosts** for your CDH cluster installation page appears.

Type in the names of five of the six hosts in the environment, separated by spaces:

`cmhost master-1 master-2 worker-1 worker-2`

Note: Do **not** include `worker-3` at this time. That host will be used in a different exercise.

Click **Search**. All five hosts should be found. Make sure that all five are selected, then click **Continue**.

<input checked="" type="checkbox"/>	Expanded Query ↑	Hostname (FQDN)	IP Address	Currently Managed	Result
<input checked="" type="checkbox"/>	cmhost	cmhost.example.com	10.0.1.19	No	Host was successfully scanned.
<input checked="" type="checkbox"/>	master-1	master-1.example.com	10.0.5.114	No	Host was successfully scanned.
<input checked="" type="checkbox"/>	master-2	master-2.example.com	10.0.2.70	No	Host was successfully scanned.
<input checked="" type="checkbox"/>	worker-1	worker-1.example.com	10.0.8.213	No	Host was successfully scanned.
<input checked="" type="checkbox"/>	worker-2	worker-2.example.com	10.0.5.207	No	Host was successfully scanned.

The **Select Repository** page appears. In this screen, you will identify the location of the CDP parcel and the Cloudera Manager Agent installer (<http://cmhost:8060/cm7.3.1/>).

- In the **Install Method** section, under **CDH and other software** area of the page, ensure the **Use parcels** option is selected.

CDH and other software

Cloudera recommends the use of parcels for installation over packages, because parcels enable Cloudera Manager to easily manage the software on your cluster, automating the deployment and upgrade of service binaries. Electing not to use parcels will require you to manually upgrade packages on all hosts in your cluster when software updates are available, and will prevent you from using Cloudera Manager's rolling upgrade capabilities.

Install Method ☐ Use Packages ?

☒ Use Parcels (Recommended) ? [Parcel Repositories & Network Settings](#)

[Other Parcel Configurations](#)

Version **Versions that are too new for this version of Cloudera Manager (7.3.1) will not be shown.**

☐ CDH 7.1.6-1.cdh7.1.6.p0.10506313

☒ CDH 7.1.5-1.cdh7.1.5.p0.7431829

Additional Parcels ☐ CFM 2.0.4.0-80

- The **Parcel Repository & Network Settings** pop-up will appear.

In the **Parcel Repository and Network Settings** popup, click on each of the garbage bin icons (🗑) to remove *ALL* the current repository references.

Click **Save & Verify Configuration**. Once you have verified all errors have been removed click on **Close** to return to the **Select Repository** page.

- Click the **Other Parcel Configurations** link

In the **Other Parcel Configurations** popup we will make a change to one of the settings. Notice that there are several properties that can be set during this installation. Change the **Parcel Update Frequency** to 5 minute(s)

Click **Save Changes** to return to the **Select Repository** page.

11. Ensure that Version Cloudera Runtime CDH-7.1.5-1.cdh7.1.5.p0.7431829 is selected. Note that there will be several versions listed. Please ensure you have the correct one selected. Although we installed Cloudera Manager 7.3.1, we will be installing 7.1.5 to the cluster. We will upgrade the cluster to 7.1.6 later in the course.

12. Click **Continue** to save the installation repository settings.

13. The **Select JDK** page appears.

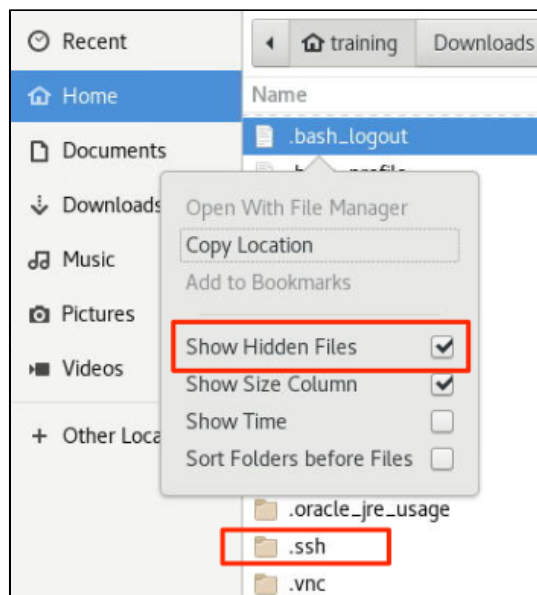
A supported version of the Oracle JDK is already installed in your exercise environment. Therefore, select the option to **Manually manage JDK**.

Click **Continue**.

14. The **Enter Login Credentials** page appears.

- Ensure that **Login To All Hosts As** is changed to **Another user** and enter **training**.
- For **Authentication Method**, choose **All hosts accept same private key**.
- Click the **Choose File** button to select a private key file.

Select **Home** in the location selector panel on the left. Then right-click in the **Name** area and select **Show Hidden Files**.

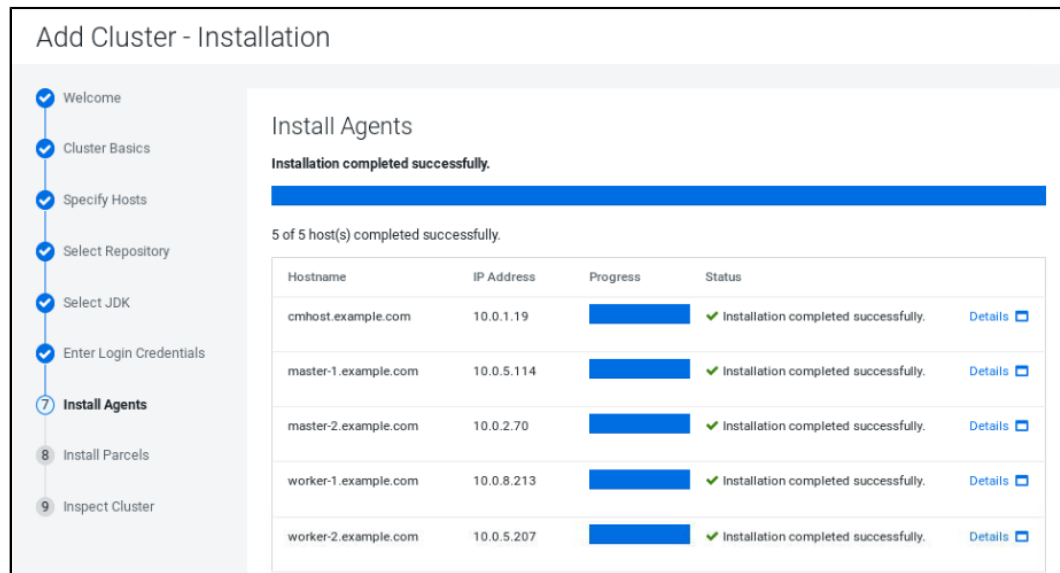


Double-click to open the **.ssh** directory, select the **id_rsa** file, and click **Open**.

d. Leave the passphrase fields blank fields and click **Continue**.

15. The **Install Agents page appears.**

Cloudera Manager installs the Cloudera Manager Agent on each machine. It may take up to five minutes.



After the Cloudera Manager Agent is successfully installed on all five machines, click **Continue** if needed.

16. In the next step, the parcel is downloaded, distributed, unpacked, and activated on all hosts in the cluster.

When the parcel is activated on all five hosts, click **Continue**.

17. The **Inspect cluster page appears.**

Select the **Inspect Network Performance** button. After a few moments, the results will appear. All validations should succeed.

Then click on the **Inspect Hosts** button. After a few moments, the results will appear. A warning that kudu is not part of group hive will appear. Click **I understand the risks of not running ...** and then click **Continue**.

If you wish you can review the reports from each inspection.

Click **Continue**.

Tip: The Host Inspector can be run on existing cluster hosts at any time from the Cloudera Manager admin console.

18. You have now completed installing Cloudera Manager and creating a cluster.

The **Add Cluster - Configuration** wizard will walk you through installing and deploying services on the cluster starts automatically.

Set up Cluster Services

19. The first page of the Add Cluster - Configuration wizard is the **Select Services** page.

Notice the note at the bottom of the screen stating “This wizard will also install the **Cloudera Management Service**.”

Click **Custom Services**, which will display a table appears with a list of CDP service types.

20. Select the following services:

- HDFS
- Hive
- Hive on Tez
- Oozie
- Spark
- Tez
- YARN
- ZooKeeper

Double check that you have selected the correct services before continuing, then click **Continue**.

21. On the next page, specify which hosts in the cluster serve which roles.

Assign roles to hosts as shown in the table below.

Note: The instructions below have you assign the HDFS DataNode role to cmhost. This is contrary to standard recommendations, even on a small cluster. Follow the instructions as shown, for now. You can modify this setup later.

Tip: To assign a role to a particular host, click on a field with one or more hostnames in it. For example, the field under **SecondaryNameNode** might initially have the value `worker-1`. Change this to a different host by clicking on the field, which will open a window where you can choose a new host. Note that the interface will include `.example.com` in all host names, as in `master-1.example.com`.

Role	Node(s)
HDFS	
NameNode	master-1
SecondaryNameNode	master-2
Balancer	master-1
HttpFS	master-1
NFS Gateway	Do not specify any hosts
DataNode	Custom: cmhost, worker-1, worker-2
Hive	
Gateway	Do not specify any hosts
Hive Metastore Server	master-1
WebHCat Server	Do not specify any hosts
HiveServer2	Do not specify any hosts
Hive on Tez	
Gateway	master-1, master-2
HiveServer2	master-2
Cloudera Management Service	
Service Monitor	cmhost
Activity Monitor	cmhost
Host Monitor	cmhost
Reports Manager	cmhost
Event Server	cmhost
Alert Publisher	cmhost
Telemetry Publisher	Do not specify any hosts
Oozie	
Oozie Server	master-2
Spark	

	Role	Node(s)
	HistoryServer	master-2
	Gateway	master-1, master-2, cmhost
Tez		
	Gateway	cmhost, master-2
YARN		
	ResourceManager	master-1
	JobHistoryServer	master-1
	NodeManager	Same as DataNode
ZooKeeper		
	Server	cmhost, master-1, master-2

When you have finished assigning roles, carefully verify that your role assignments are correct. When you are certain that the settings are correct, click **Continue**.

- 22.** The **Setup Database** page appears. This allows you to specify the database connection details for each service's database in the Cloudera Manager database system. The databases and access credentials were created when you ran `scm_prepare_database.sh` during the Cloudera Manager installation process.

Note: The Database Hostname should be set to `cmhost` or `cmhost.example.com` for each service. Either one will work.

Fill in the details as shown here.

Service	Database Hostname	Database Type	Database Name	Username	Password
Hive	cmhost	MySQL	metastore	hiveuser	password
Activity Monitor	cmhost	MySQL	amon	amonuser	password
Reports Manager	cmhost	MySQL	rman	rmanuser	password
Oozie Server	cmhost	MySQL	oozie	oozieuser	password

Click **Test Connection** to verify that Cloudera Manager can connect to the MySQL databases you created in an earlier exercise in this course.

After you have verified that all connections are successful, click **Continue**.

23. The **Review Changes** page appears. Leave the default values for all settings. Click **Continue**.
-

24. The **First Run Command** page appears. Progress messages appear while cluster services are created and started. When all the cluster services have started, click **Continue**.
-

25. The **Summary** page appears.

The page indicates that services have been added and are now configured and running on your cluster. Click **Finish**.

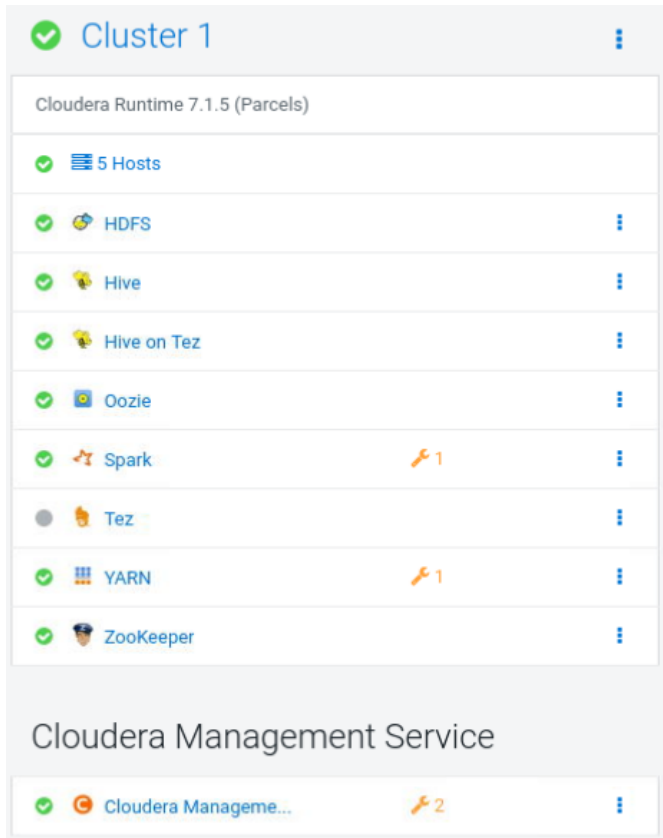
The Cloudera Manager home page will appear.

26. On the home page, confirm that all services are running correctly—that is, they should have a green checkmark indicator. If any of the services are not running, open the drop-down menu to the right of the service name and select **Restart**.

Configuration Warnings

The configuration warnings—as seen in the screenshot below—are expected, and indicate that, although Cloudera Management Services and the Service Monitor are in good health, but they do not have the recommended amount of memory available. There is also an indicator that YARN needs the YARN Queue Manager which we will install later. There is also an erroneous message indicating that we need a Spark Gateway on master-1 and master-2, which already exist.

In a production deployment, you would need to ensure these warnings were addressed. However, in the exercise environment, you can safely ignore them at this time.



Cluster installation is now complete.

This is the end of the exercise.

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.

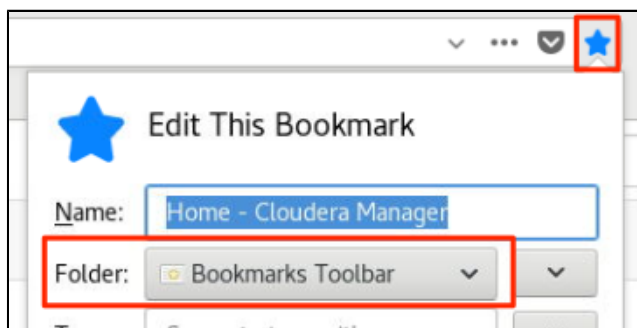
Modify a service configuration

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.
-

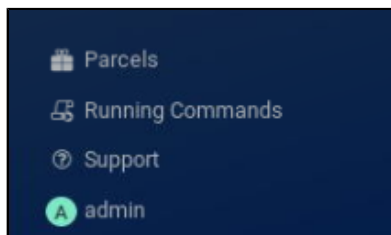
13. *Optional:* Enter a description of the change, such as `scan every 10 days`.
-

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`

NiFi Node Advanced Configuration Snippet (Safety Valve) for staging/state-management.xml

View as XML

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

Value: `true`

Description:

☐ Final

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

Value: `true`

Description:

☐ Final

- 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.