HITACHI
Inspire the Next

Getting Started with Pentaho and Cloudera QuickStart VM

This page intentionally left blank.

Contents

This page intentionally left blank.

Overview

This document covers some best practices on integrating Pentaho software with Cloudera QuickStart Virtual Machine (VM). In it, you will learn how to configure the QuickStart VM so that Pentaho can connect to it.

Our intended audience is Pentaho developers and system architects looking to experiment with Pentaho Data Integration (DI) and Hadoop.

The intention of this document is to speak about topics generally; however, these are the specific versions covered here:

Software	Version(s)
Pentaho	7.x, 8.x

The Components Reference in Pentaho Documentation has a complete list of supported software and hardware.

Before You Begin

Before beginning, use the following information to prepare for the procedures described in the main section of the document.

Prerequisites

This document assumes that you have knowledge about Pentaho Data Integration (PDI) and have installed Pentaho software.

Note that this guide was developed with VMWare. If you are using other platforms like VirtualBox or KVM, you will need to adapt some of the concepts described in this document to your particular platform.

Use Case: Development Sandbox for Pentaho and Hadoop

> *Janice is a Pentaho administrator setting up a new installation. She wants to make sure that her configuration works perfectly before she implements anything in production, so she has decided to set up a development sandbox where she can test and experiment with the integration between Pentaho and Hadoop.*

Before Configuring Cloudera QuickStart VM

There are a few tasks that must be performed to correctly configure the Cloudera QuickStart VM so that Pentaho can connect to it.

Note that this guide was developed with VMWare. If you are using other platforms like VirtualBox or KVM, you will need to adapt some of the concepts described in this document to your particular platform.

You can find details on these topics in the following sections:

- Download Cloudera QuickStart VM
- Upgrade VM (VMWare Version)
- Set Memory, CPU, and Networking Settings
- Increase Filesystem Size

Download Cloudera QuickStart VM

To download Cloudera QuickStart VM:

1. Visit Cloudera's QuickStarts for CDH.
2. Select the **VMWare** platform.
3. Click the **GET IT NOW** button.
4. Extract the downloaded package of the Cloudera QuickStart VM.

Upgrade VM (VMWare Version)

Depending on the version of VMWare that you are using, the Cloudera QuickStart VM might be outdated. Update the VM to the latest version of VMWare to get all its benefits.

To perform this upgrade, open VMWare:

1. Open the VM and click the **Upgrade this virtual machine** button:

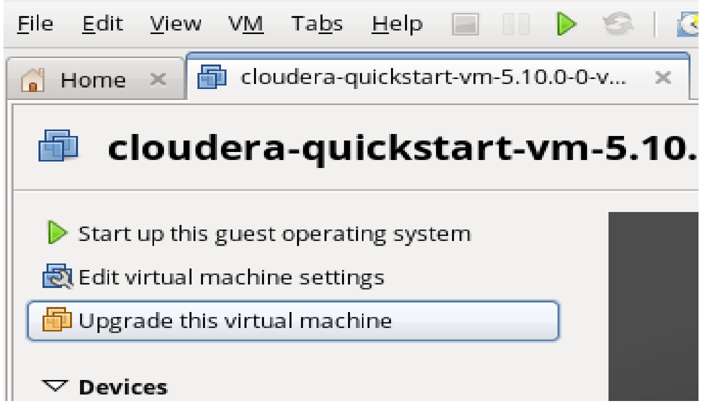
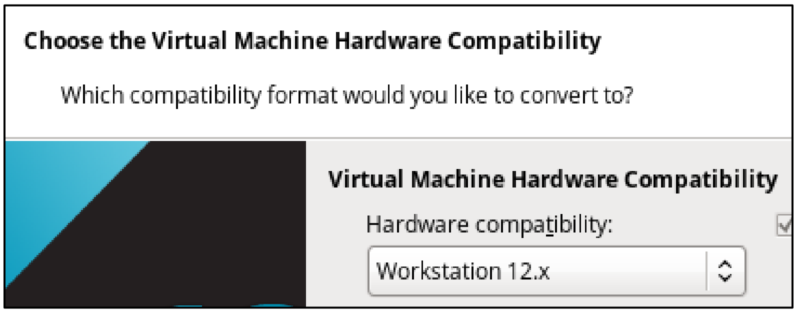

File
Edit
View
VM
Tabs
Help
Home
cloudera-quickstart-vm-5.10.0-0-v...
cloudera-quickstart-vm-5.10.
Start up this guest operating system
Edit virtual machine settings
Upgrade this virtual machine
Devices

Figure 1: Upgrade VM

2. Select your VMWare version in the **Virtual Machine Hardware Compatibility** dropdown and click **Next**:



Choose the Virtual Machine Hardware Compatibility

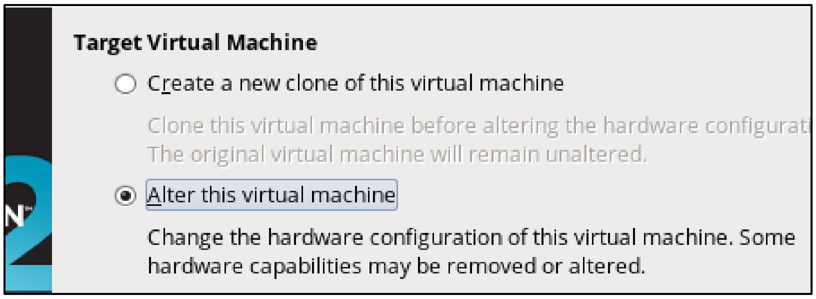
Which compatibility format would you like to convert to?

Virtual Machine Hardware Compatibility

Hardware compatibility: Workstation 12.x

Figure 2: Select VMWare Version

3. Select **Alter this virtual machine** in the **Target Virtual Machine** window:



Target Virtual Machine

○ Create a new clone of this virtual machine

Clone this virtual machine before altering the hardware configurati
The original virtual machine will remain unaltered.

● Alter this virtual machine

Change the hardware configuration of this virtual machine. Some hardware capabilities may be removed or altered.

Figure 3: Alter This Virtual Machine

Set Memory, CPU, and Networking Settings

Verify the following settings and qualifications in the VMWare:

- Use at least 16GB of RAM.
- Use at least two CPUs (or cores).
- Change the **Network** to be **Host Only**.

Increase Filesystem Size

The Cloudera QuickStart VM is configured with a 64GB volume. While this is enough to run Cloudera's basic functionality, 64GB is the minimum amount of space and can cause out of space errors.

To increase the available space, you will need to add a new volume using VMWare's VM settings, and then expand the logical volume using Logical Volume Management (LVM):

Add a New Volume

Add the new volume to the VM using these steps:

1. Open a terminal and gain superuser (`root`) access using the `sudo su` command:

```
[cloudera@quickstart ~]$ sudo su -
[root@quickstart ~]#
```

2. With the superuser terminal, confirm the new volume is correct attached by running `fdisk -l /dev/sdb`:

```
[root@quickstart ~]# fdisk -l /dev/sdb

Disk /dev/sdb: 107.4 GB, 107374182400 bytes
255 heads, 63 sectors/track, 13054 cylinders
Units = cylinders of 16065 * 512 = 8225280 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x00000000
```

3. Create the necessary partition in the new volume `/dev/sdb` using `fdisk`:

```
[root@quickstart ~]# fdisk /dev/sdb
```

4. Type n to create a new partition:

```
Command (m for help): n
```

5. When you are prompted, enter `p` for a primary partition, and then `1` for the partition number:

```
Command action
   e   extended
   p   primary partition (1-4)
p
Partition number (1-4): 1
```

6. Use **Enter** to accept the default values for the all the remaining questions:

```
First cylinder (1-13054, default 1):
Using default value 1
Last cylinder, +cylinders or +size{K,M,G} (1-13054, default 13054):
Using default value 13054

Command (m for help): t
```

```
Selected partition 1
Hex code (type L to list codes): 8e
Changed system type of partition 1 to 8e (Linux LVM)

Command (m for help): w
The partition table has been altered!

Calling ioctl() to re-read partition table.
Syncing disks.
```

Increase the Logical Volume

Next, you will need to increase the logical volume.

1. First, use `pvcreate` to create the physical volume for LVM to use, based on the new partition `/dev/sdb1`:

```
[root@quickstart ~]# pvcreate /dev/sdb1
  Physical volume "/dev/sdb1" successfully created
```

2. Extend the volume group by adding the new physical volume:
   a. Confirm the name of the volume group using `vgdisplay`:

```
[root@quickstart ~]# vgdisplay
  --- Volume group ---
  VG Name               vg_quickstart
```

   b. Extend the volume group using `vgextend`:

```
[root@quickstart ~]# vgextend vg_quickstart /dev/sdb1
  Volume group "vg_quickstart" successfully extended
```

   c. Scan the physical volumes with `pvscan` to confirm which volumes are used by the volume group:

```
  PV /dev/sda2   VG vg_quickstart   lvm2 [63.51 GiB / 0    free]
  PV /dev/sdb1   VG vg_quickstart   lvm2 [100.00 GiB / 100.00 GiB free]
  Total: 2 [163.50 GiB] / in use: 2 [163.50 GiB] / in no VG: 0 [0   ]
```

3. Increase the logical volume:
   a. Confirm the path of the logical volume with `lvdisplay`:

```
[root@quickstart ~]# lvdisplay
  --- Logical volume ---
  LV Path                /dev/vg_quickstart/lv_root
```

b. Extend the volume using `lvextend`:

```
[root@quickstart ~]# lvextend /dev/vg_quickstart/lv_root  /dev/sdb1

  Size of logical volume vg_quickstart/lv_root changed from 55.51 GiB
(14210 extents) to 155.50 GiB (39809 extents).

  Logical volume lv_root successfully resized
```

c. Run `vgdisplay` and `lvdisplay` to confirm the volume group (`VG Size`) and logical volume (`LV Size`) sizes.

4. Resize the filesystem using `resize2fs`:

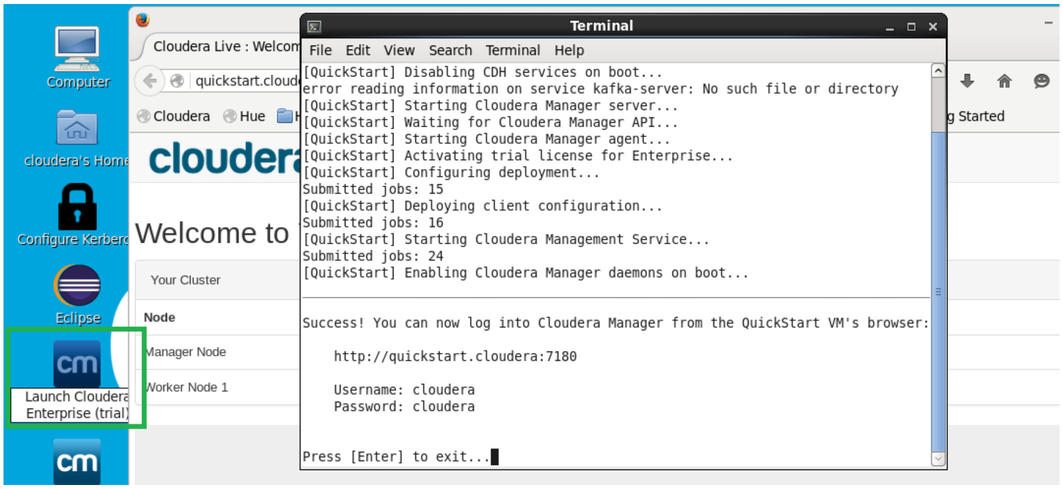
```
[root@quickstart ~]# resize2fs /dev/vg_quickstart/lv_root

resize2fs 1.41.12 (17-May-2010)

Filesystem at /dev/vg_quickstart/lv_root is mounted on /; on-line resizing
required

old desc_blocks = 4, new_desc_blocks = 10

Performing an on-line resize of /dev/vg_quickstart/lv_root to 40764416 (4k)
blocks.

The filesystem on /dev/vg_quickstart/lv_root is now 40764416 blocks long.
```

5. Confirm the filesystem's new size using the command `df -h`:

```
[root@quickstart ~]# df -h

Filesystem            Size  Used Avail Use% Mounted on

/dev/mapper/vg_quickstart-lv_root

                      153G  8.2G  138G   6% /
```

Start and Configure Cloudera Enterprise (Trial)

To start Cloudera Enterprise, use the **Launch Cloudera Enterprise (trial)** icon on the desktop of the Cloudera QuickStart VM, and wait for the terminal window to finish:



```
[QuickStart] Disabling CDH services on boot...
error reading information on service kafka-server: No such file or directory
[QuickStart] Starting Cloudera Manager server...
[QuickStart] Waiting for Cloudera Manager API...
[QuickStart] Starting Cloudera Manager agent...
[QuickStart] Activating trial license for Enterprise...
[QuickStart] Configuring deployment...
Submitted jobs: 15
[QuickStart] Deploying client configuration...
Submitted jobs: 16
[QuickStart] Starting Cloudera Management Service...
Submitted jobs: 24
[QuickStart] Enabling Cloudera Manager daemons on boot...

Success! You can now log into Cloudera Manager from the QuickStart VM's browser:

    http://quickstart.cloudera:7180

    Username: cloudera
    Password: cloudera


Press [Enter] to exit...
```

Figure 4: Starting Cloudera Enterprise

Increase Memory for Services

You will need to configure both Cloudera and Hadoop Distributed Filesystem (HDFS) components' Java heap settings to avoid out of memory errors. First, ensure you have allocated at least 16GB of RAM to the Cloudera VM.

To configure memory settings:

1. Open the **Cloudera Manager** and go to the **Home** screen.
2. Click both **Cloudera Manager** and **HDFS**.
3. Click the **Configuration** menu and type **Java Heap** in the **Search** input.

4. Change each of the following properties:

Table 1: Memory Settings Changes

Item	Setting	Change From	Change To
Cloudera Manager	Java heap size of Host Monitor	256MB	1GB
	Maximum non-Java memory of Host Monitor	768MB	1.5GB
	Java heap size of Service Monitor	256MB	1GB
	Maximum non-Java memory of Service Monitor	768MB	1.5GB
HDFS	Java heap size of `NameNode`	50MB	1GB
	Java heap size of secondary `NameNode`	50MB	1GB

5. Click **Save Changes**.
6. Exit the **Configuration** screen.
7. If any services were already started, restart them through the **Home** screen in Cloudera Manager.
8. Start all other required services through Cloudera Manager's **Home** screen as well by clicking the dropdown menu to the right of the title and selecting **Start**.

PDI Configuration

There are a few items to configure before PDI can communicate with the Cloudera VM.

You can find details on these topics in the following sections:

- Before Configuring PDI
- Configuring PDI
- Connecting to Cloudera QuickStart VM
- Testing PDI Functionality

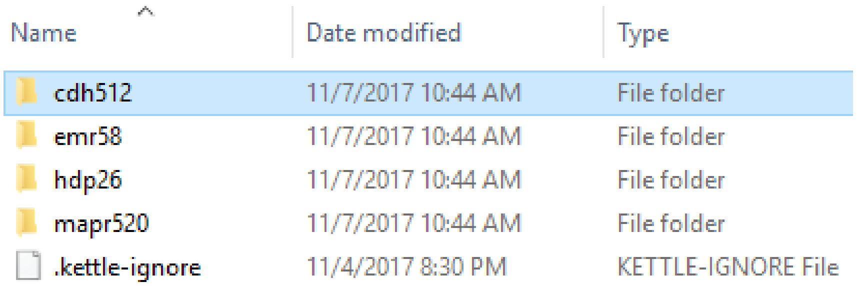
Before Configuring PDI

Before you can configure PDI, make sure you have installed PDI along with the correct shim. More information is available at Components Reference – Big Data Sources.

Shim

The shim, a set of libraries and configurations required to connect to each Big Data distribution, isolates the development of the PDI jobs and transformations from the specifics for each distribution. Depending on your PDI and Cloudera versions, you may or may not need a shim.

In your installed PDI folder structure, look in your `data-integration/plugins/pentaho-big-data-plugin/hadoop-configurations` folder and identify the required Cloudera version. For example, this folder shows a CDH version of 5.12:



Name	Date modified	Type
cdh512	11/7/2017 10:44 AM	File folder
emr58	11/7/2017 10:44 AM	File folder
hdp26	11/7/2017 10:44 AM	File folder
mapr520	11/7/2017 10:44 AM	File folder
.kettle-ignore	11/4/2017 8:30 PM	KETTLE-IGNORE File

Figure 5: Hadoop Configurations Folder Contents

If your installed Cloudera version matches the version indicated on your folder, you do not need a shim.

If it does not match, you can look up and download the appropriate shim on the Downloads page of the Pentaho Customer Portal.

To configure your installation for your shim:

1. Edit the `data-integration/plugins/pentaho-big-data-plugin/plugin.properties` file.
2. Make sure the `active.hadoop.configuration` property is pointing to the correct shim. For example, the property value must be `cdh510` for Cloudera QuickStart VM version 5.10.

Configuring PDI

To configure PDI, download the Cloudera configuration files from Cloudera Manager.

1. Access your Cloudera Manager (`http://localhost:7180`).
2. In the components list, below the cluster name (Cloudera QuickStart), click the **Hive** component.
3. On the Hive page in Cloudera Manager, click **Actions** in the drop-down menu and then select **Download Client Configuration**.
4. Extract the configuration XML files `core-site.xml`, `hive-site.xml`, `mapred-site.xml`, and `yarn-site.xml` into your CDH folder (for Cloudera QuickStart 5.10, that folder is `data-integration/plugins/pentaho-big-data-plugin/hadoop-configurations/cdh510`).
5. Restart Spoon or any other PDI component that will use this configuration.

For a more detailed guide, visit Set Up Pentaho to Connect to a Cloudera Cluster.

Connecting to Cloudera QuickStart VM

To connect to the Cloudera QuickStart VM:

1. In Spoon, create a new job or transformation.
2. Click on the **View** tab.
3. Right-click the **Hadoop clusters** entry and select **New cluster**:

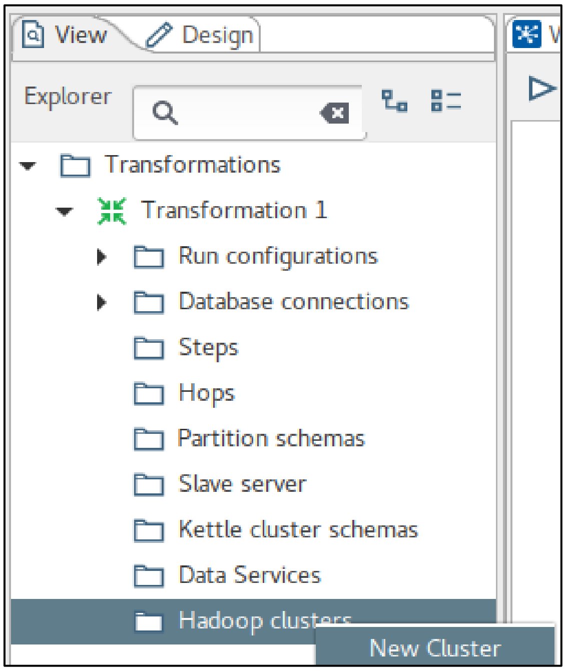
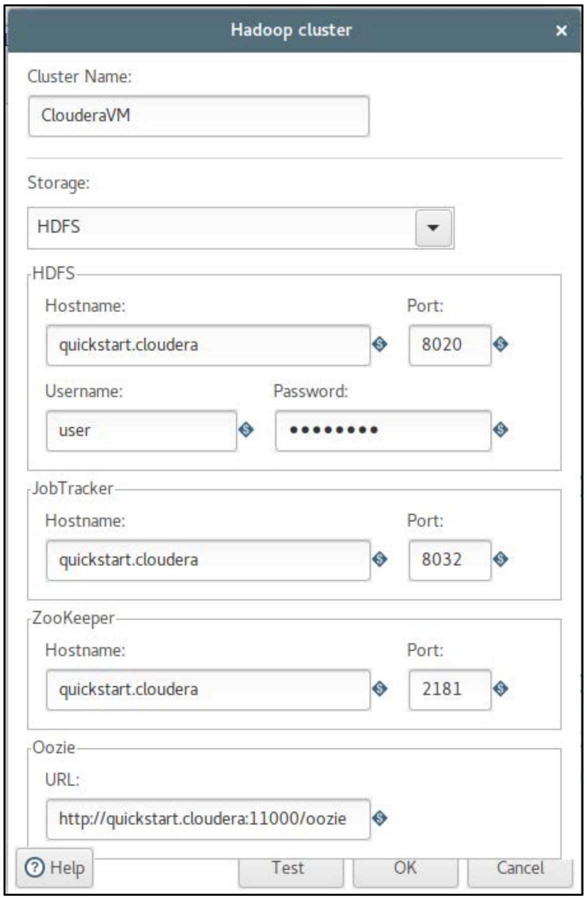


Figure 6: New Cluster

4. In the new **Hadoop cluster** screen, assign a name to your cluster in the **Cluster Name** text box.

5. Select **HDFS** as the storage.
6. Enter the correct **host** and **port** for each component (Use the `quickstart.cloudera` as host. If you are not connecting through the VM, then create a `hosts` file to resolve `quickstart.cloudera` to the host IP of the VM).



Cluster Name: ClouderaVM

Storage: HDFS

HDFS
Hostname: quickstart.cloudera
Port: 8020
Username: user
Password: •••••••••

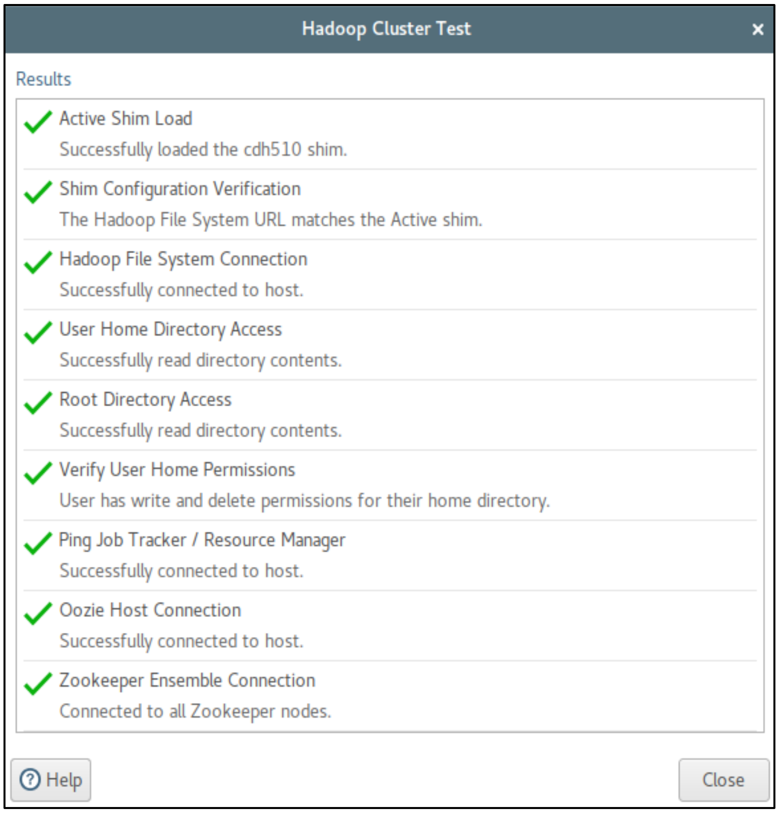
JobTracker
Hostname: quickstart.cloudera
Port: 8032

ZooKeeper
Hostname: quickstart.cloudera
Port: 2181

Oozie
URL: http://quickstart.cloudera:11000/oozie

Figure 7: Hadoop Cluster Properties

7. At the bottom of the Hadoop cluster screen, click the **Test** button. Make sure everything shows with a green check mark. If it does, click **Close** and then **OK**, and save your transformation.



✓ Active Shim Load
Successfully loaded the cdh510 shim.

✓ Shim Configuration Verification
The Hadoop File System URL matches the Active shim.

✓ Hadoop File System Connection
Successfully connected to host.

✓ User Home Directory Access
Successfully read directory contents.

✓ Root Directory Access
Successfully read directory contents.

✓ Verify User Home Permissions
User has write and delete permissions for their home directory.

✓ Ping Job Tracker / Resource Manager
Successfully connected to host.

✓ Oozie Host Connection
Successfully connected to host.

✓ Zookeeper Ensemble Connection
Connected to all Zookeeper nodes.

Figure 8: Hadoop Cluster Test

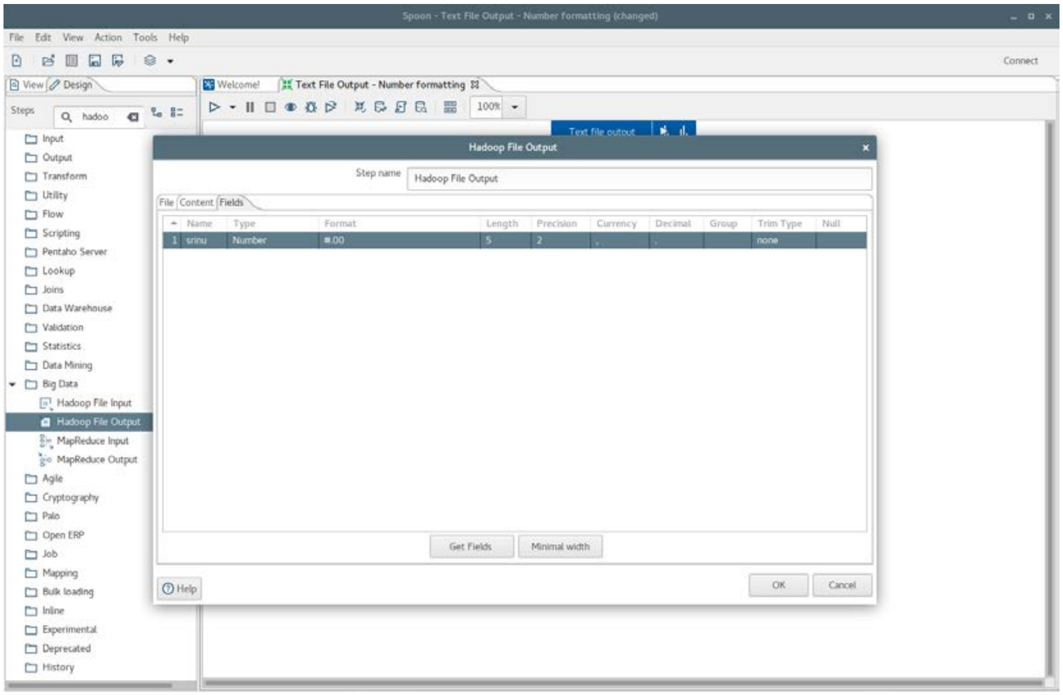
8. If you receive an error on the **User Home Directory Access** test (one of the tests in Hadoop Cluster Test), you will need to create a directory in HDFS for the user who is running Spoon:

```
[cloudera@quickstart ~]$ sudo su - hdfs
-bash-4.1$ hdfs dfs -mkdir /user/rodrigo
-bash-4.1$ hdfs dfs -chmod 777 /user/rodrigo
-bash-4.1$
```

Testing PDI Functionality

An easy way to test PDI functionality is to use Spoon to output data into HDFS using the **Hadoop File Output** step. You can create your own transformation from scratch.

For a faster test, you can use the `Text File Output - Number formatting.ktr` file located in your `data-integration/samples/transformations/` folder and add the **Hadoop File Output** step at the end.

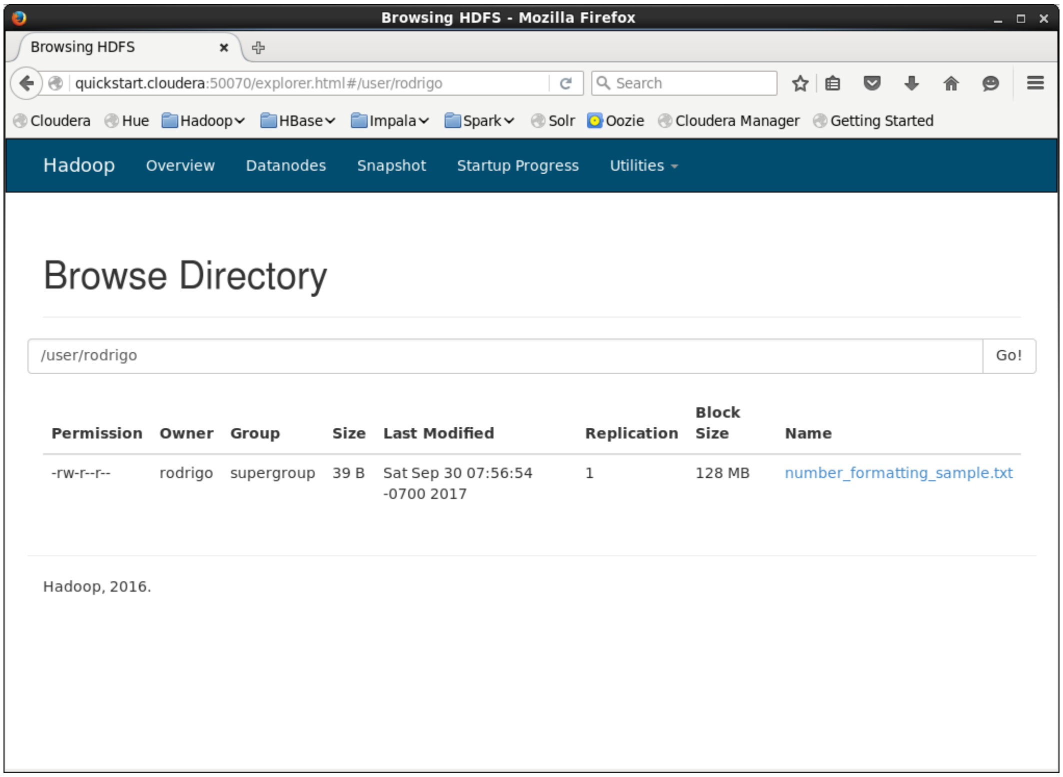


Step name: Hadoop File Output

	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim Type	Null
1	srinu	Number	#.00	5	2	,	.		none	

Figure 9: Hadoop File Output Fields

It is also possible to confirm that the file exists by browsing the filesystem with the NameNode Web User Interface:



Browse Directory

/user/rodrigo

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	rodrigo	supergroup	39 B	Sat Sep 30 07:56:54 -0700 2017	1	128 MB	number_formatting_sample.txt

Hadoop, 2016.

Figure 10: NameNode

Or use the `HDFS DFS` command to explore the filesystem:

```
-bash-4.1$ hdfs dfs -ls /user/rodrigo/
Found 1 items
-rw-r--r--   1 rodrigo supergroup         39 2017-09-30 07:56
/user/rodrigo/number_formatting_sample.txt
-bash-4.1$
```

Troubleshooting

Here are some possible errors you might receive, and how to resolve them:

Requests to Service Monitor and Host Monitor Failed

If you receive an error that both requests to the Service Monitor and Host Monitor failed, restart the Cloudera Manager service through the Cloudera Manager Home page.

Hadoop File System URL Does Not Match

When testing the Hadoop cluster entry in Spoon, if you receive the warning "The Hadoop File System URL does not match the URL in the `shim core-site.xml`" in the "Shim Configuration Verification" test, please make sure that you are using `quickstart.cloudera` as the host.

Related Information

Here are some links to information that you may find helpful while using this best practices document:

- Download QuickStarts for CDH 5.12
- Pentaho Components Reference
- Pentaho Components Reference – Big Data Sources
- Pentaho Customer Portal – Downloads
- Set Up Pentaho to Connect to a Cloudera Cluster

Finalization Checklist

This checklist is designed to be added to any implemented project that uses this collection of best practices, to verify that all items have been considered and reviews have been performed.

Name of the Project:____

Date of the Review:____

Name of the Reviewer:____

Item	Response	Comments
Did you download Cloudera QuickStart VM?	YES________ NO________	
Did you upgrade your VMWare?	YES________ NO________	
Did you configure your memory, CPU, and network settings?	YES________ NO________	
Did you increase your filesystem size?	YES________ NO________	
Did you increase memory for your Java heap settings?	YES________ NO________	
Did you install your shim, if you needed one?	YES________ NO________	
Did you configure PDI?	YES________ NO________	
Did you connect PDI to Cloudera QuickStart VM?	YES________ NO________	
Did you test PDI's functionality?	YES________ NO________	