HITACHI
Inspire the Next

Configuring PDI, Pentaho MapReduce, and MapR

HITACHI
Inspire the Next

Change log (if you want to use it):

Date	Version	Author	Changes

Contents

This page intentionally left blank.

Overview

This document is intended to provide insight and best practices for setting up Pentaho Data Integration (PDI) to work with MapR. It includes information about setting up and installing the MapR client tool that is required by PDI to run Pentaho MapReduce (PMR) jobs.

Certain configurations on the Hadoop ecosystem will be examined to make sure that the client is correctly setup before PDI will use it.

The Components Reference in Pentaho Documentation has a complete list of software versions for compatibility between your Pentaho and JDK versions..

The information in this document covers the following versions:

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, make sure that you are using a MapR client that is compatible with Windows OS.

Set Up Your Environment

This section covers the things you need to do to set up your environment.

You can find details on these topics here:

- Getting MapR Server Information
- Setting Up Your Host Environment
- Installing the Pentaho Shim for MapR

Getting MapR Server Information

We have used a MapR Virtual Machine (VM) downloaded from MapR Sandbox to show you how to configure PDI to work with MapR. We are also using a Windows development environment for this demonstration.

Add a `host`-only adaptor to the VM once it has been imported into either VMware or VirtualBox. You will be able to obtain the IP address of the VM once it starts.

The IP address may vary depending on your environment. This IP will be used throughout this process

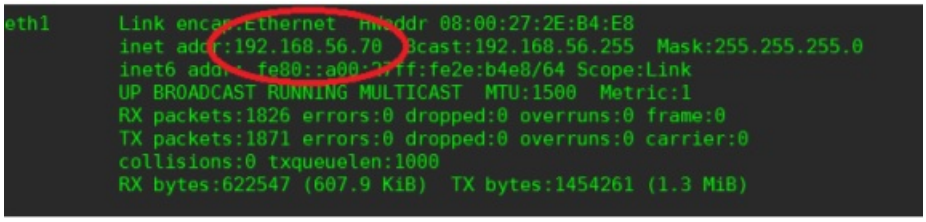

Figure 1: Windows Development Environment

Setting Up Your Host Environment

You will need to update your `hosts` file in your development environment, which could be either Windows, MacOS, or Unix, so that it points to this virtual machine. The location of the `hosts` file is `192.168.56.70`, as shown:

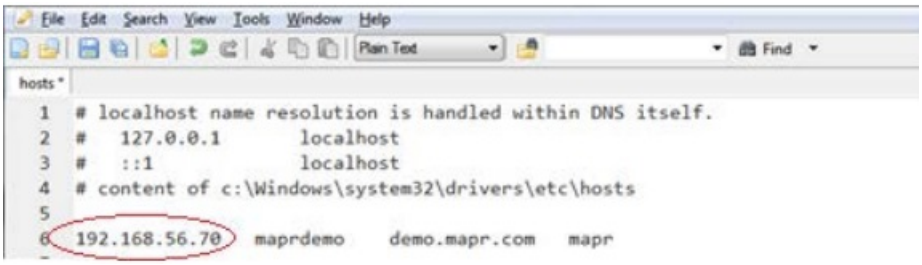

Figure 2: Host Environment

You will now be able to use your browser to connect to the MapR administration page using either of the following URLs:

```
http:/demo.mapr.com:8443/mcs
or
http:/localhost:8443/mcs
```

You will be able to log in using the credentials `root/mapr` or `mapr/mapr` to get to the main console:

The cluster name is `demo.mapr.com`. This cluster name will be required when you configure the MapR client.

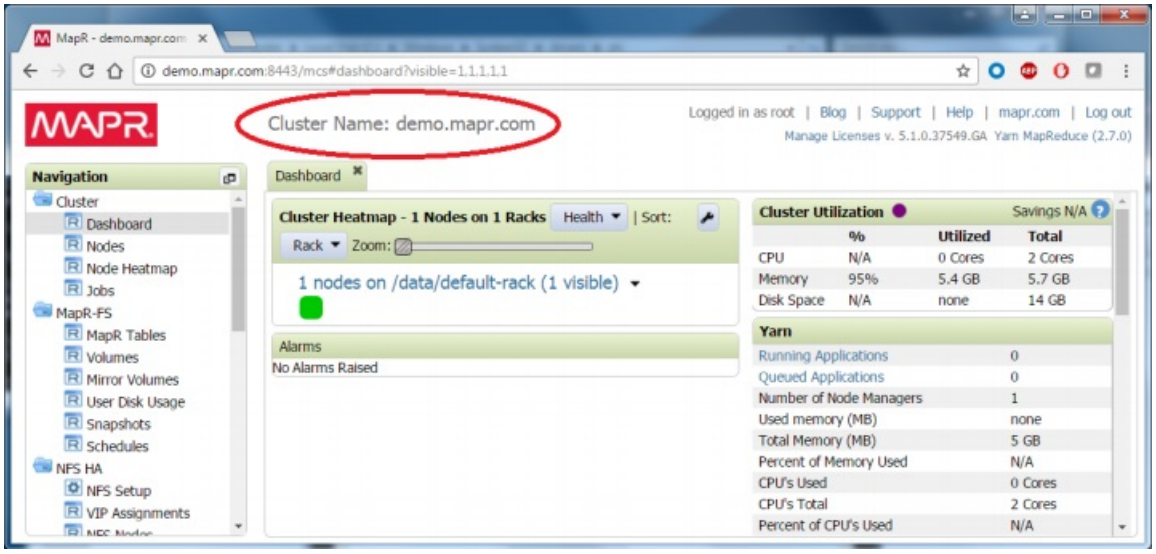

Figure 3: MapR Cluster

Installing the Pentaho Shim for MapR

You will need to install the Pentaho Shim for MapR that comes with the software.

The shim files `core-site.xml`, `mapred-site.xml`, and `hdfs-site.xml` are not required to be installed on the Pentaho shim folder because we are using the MAPR client software.

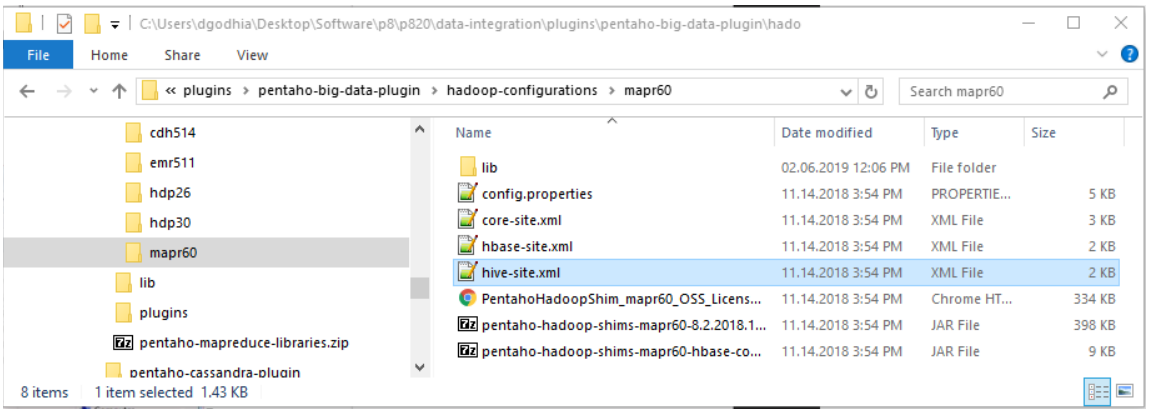

Figure 4: Pentaho Shim for MapR

Install and Configure MapR Client

This section will help with downloading and configuring the MapR client tools, as well as setting up the environment variables.

Downloading MapR Client Tools

The latest MapR client tools can be downloaded from the index. The index contains client tools for various operating systems. We are installing the client tool onto a Windows operating system for this demonstration.

We recommend extracting the downloaded client tool (zip file format) into `C:\opt\mapr`.

The folder structure of the client will show the following directories:

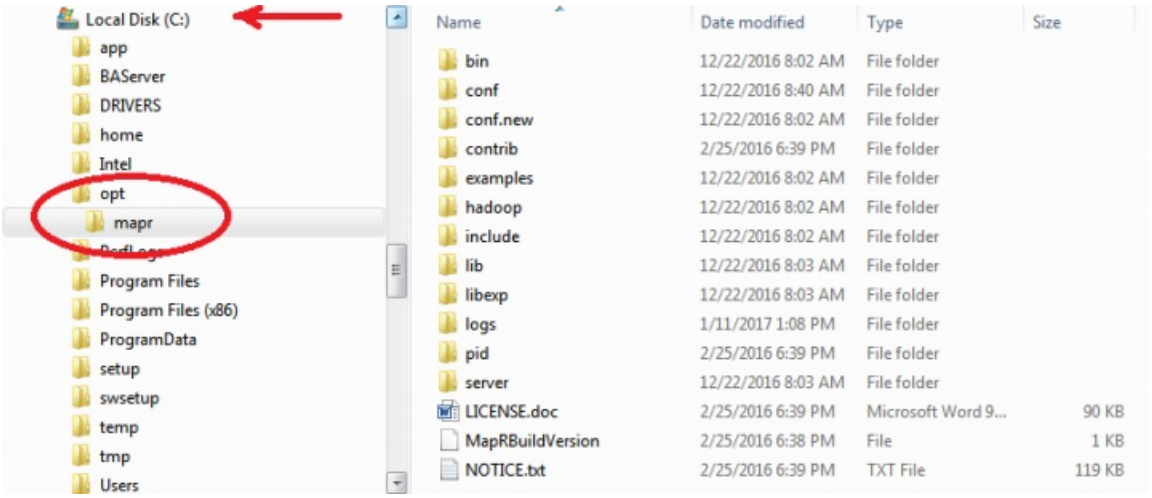

Figure 5: MapR Client Installation

Setting Up the Environment Variables

The MapR client requires some configuration changes that allow you to connect to the Hadoop filesystem using the client tools. You will be able to do this once it is installed. MapR's documentation has specific information about the ports used by MapR.

1. Open the command prompt and follow the instructions to configure the client:

```
cd c:\opt\mapr\server
set MAPR_HOME=c:\opt\mapr
```

```
configure.bat -N demo.mapr.com -c -C maprdemo:7222 -HS maprdemo
```

2. The values assigned to each of the parameters are indicated below:

```
-N cluster name (obtained earlier from the browser screen)
-C nodes running CLDB services and port
-HS history server name – host name found in the hosts file
```

3. You may also need to set the JAVA_HOME variable if this is not already set in your environment. So that PDI can connect to the MapR environment, the variable MAPR_HOME needs to be set globally in a Windows environment, and the variables `.bash_profile` or `.bash_rc` are needed globally in a Unix environment:

```
set MAPR_HOME=c:\opt\mapr
```

Use `set-pentaho-env.bat` to set the MAPR_HOME environment variable in Windows. For Unix, use `set-pentaho-env.sh`.

Configuring MapR Client to Connect into HDFS

There are a couple of files you will need to configure in the MapR Client. You will need the user ID for these files. After you have finished configuring them, you will be able to connect to HDFS. This is explained in the following sections:

Modifying core-site.xml for MapR Client

Add these lines to the following file for the MapR client's `core-site.xml`, found in the `c:\opt\mapr\hadoop\hadoop-x.x.x\etc\hadoop\` directory:

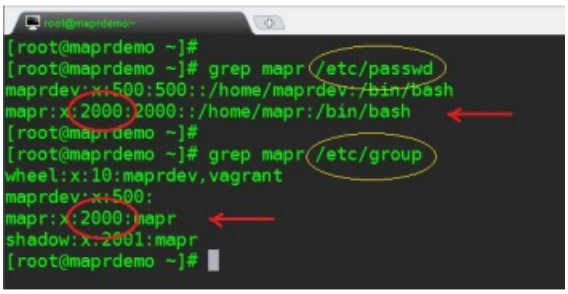

Figure 6: MapR VM

```
<property>
 <name>hbase.table.namespace.mappings</name>
 <value>*:/tables</value>
```

```
</property>
<property>
 <name>hadoop.proxyuser.mapr.hosts</name>
 <value>*</value>
</property>
<property>
 <name>hadoop.proxyuser.mapr.groups</name>
 <value>*</value>
</property>
<property>
 <name>hadoop.spoofed.user.uid</name>
 <value>2000</value>
</property>
<property>
 <name>hadoop.spoofed.user.gid</name>
 <value>2000</value>
</property>
<property>
 <name>hadoop.spoofed.user.username</name>
 <value>mapr</value>
</property>
```

The user ID (`uid`) value in `hadoop.spoofed.user.uid` can be obtained from the `/etc/passwd` file found in the MapR VM for the Hadoop user.

The group ID (`gid`) value in `hadoop.spoofed.user.gid` can be obtained from the `/etc/group` file found in the MapR VM for the user that is going to be used as the Hadoop user.

Modifying mapred-site.xml for MapR Client

A cross-platform parameter needs to be added to the `mapred-site.xml` in order to run PDI in the Hadoop cluster.

The `mapred-site.xml` file is found in the `C:\opt\mapr\hadoop\hadoop-x.x.x\etc\hadoop` directory:

```
<property>
 <name>mapreduce.app-submission.cross-platform</name>
 <value>true</value>
</property>
```

Connecting into HDFS Using MapR Client

You will be able to test the MapR client once your environment has been configured.

The configurations listed above are the bare minimum to test the connectivity and run PDI jobs in MapR.

1. Open the command window and navigate to the `C:\opt\mapr\hadoop\hadoop-x.x.x\bin` directory.
2. Make sure that the `MAPR_HOME` environment variable has previously been set.
   If the setup is incorrect you may see `UID_2000:GID_2000` instead of `mapr:root`, as shown below in the ownership of HDFS' directories.

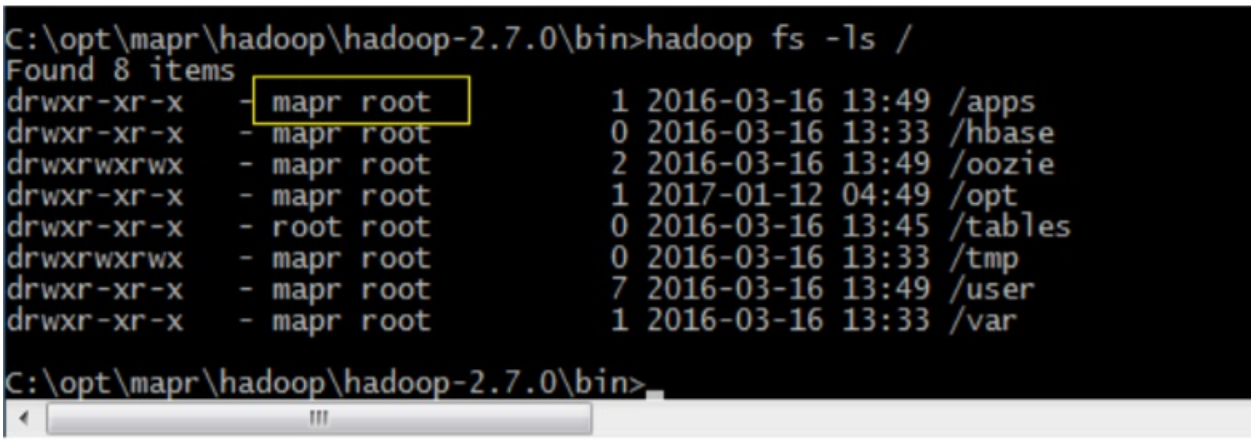

Figure 7: MapR for HDFS

```
cd C:\opt\mapr\hadoop\hadoop-2.7.0\bin
set MAPR_HOME
hadoop fs -ls /
C:\opt\mapr\hadoop\hadoop-0.20.2\bin>hadoop fs -ls /
```

Configure Hadoop Cluster Environment for PDI Jobs

The last step in the configuration is to verify that the libraries for the MapR client are loaded when PDI MapReduce is executed.

- Selecting the Hadoop Distribution for MapR
- Modifying `config.properties` in Pentaho Shim Folder
- Running PDI PMR for Samples

Selecting the Hadoop Distribution for MapR

From the **Tools** menu of the PDI, select the **Hadoop Distribution**. You will be presented with a selection box. Make sure that **MapR** is your selected Hadoop environment as shown below:

*Restart the PDI client after you select **MapR** for your Hadoop distribution.*

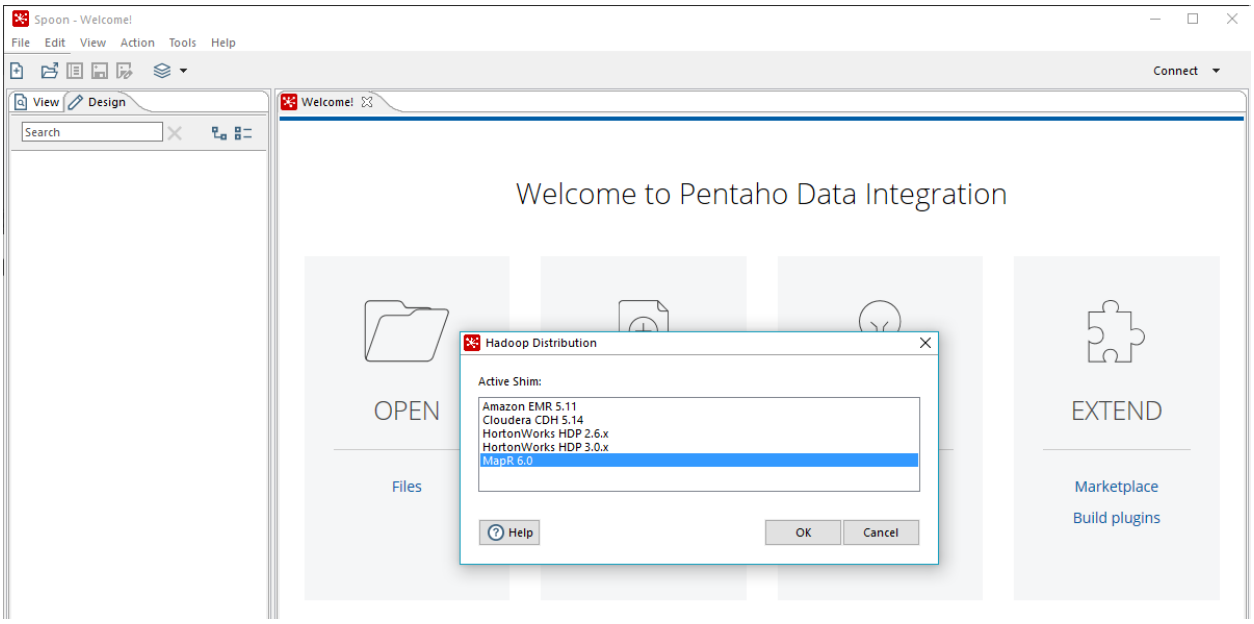

Figure 8: Hadoop Distribution for MapR

Modifying config.properties in Pentaho Shim Folder

In this example, we are using version 5.1, which corresponds with the MapR VM. You will need to change some of the locations and code based on your version. The file can be found in:

```
<pentaho_folder>\data-integration\plugins\pentaho-big-dataplugin\hadoop-configurations\mapr510
```

1. Edit the `config.properties` found in the Pentaho shim folder.
2. Edit the following values. Make sure to keep the triple dashes and change all the `hadoop2.7.0` and `mapr510` references to your correct version:

```
windows.classpath=lib/hadoop2-windows-
patch08072014.jar,file:///C:/opt/mapr/hadoop/hadoop2.7.0/etc/hadoop,file://
/C:/opt/mapr/hadoop/hadoop2.7.0/share/hadoop/common/lib,file:///C:/opt/mapr
/hadoop/hadoop2.7.0/share/hadoop/common,file:///C://opt/mapr/hadoop/hadoop2
.7.0/share/hadoop/hdfs,file:///C:/opt/mapr/hadoop/hadoop2.7.0/share/hadoop/
hdfs/lib,file:///C:/opt/mapr/hadoop/hadoop2.7.0/share/hadoop/yarn/lib,file:
///C:/opt/mapr/hadoop/hadoop2.7.0/share/hadoop/yarn,file:///C:/opt/mapr/had
oop/hadoop2.7.0/share/hadoop/mapreduce/lib,file:///C:/opt/mapr/hadoop/hadoo
p2.7.0/share/hadoop/mapreduce,file:///C:/Pentaho/design-
tools/dataintegration/plugins/pentaho-big-data-
plugin/hadoopconfigurations/mapr510,file:///C:/Pentaho/design-
tools/dataintegration/plugins/pentaho-big-data-
plugin/hadoopconfigurations/mapr510/lib,file:///C:/opt/mapr/lib

windows.library.path=C:///opt///mapr///lib
```

The `windows.classpath` and `windows.library.path` values are dependent on the version of the MapR client that you installed.

1. Start or restart the PDI tool by running `spoon.bat` or `spoon.sh`.
2. Create a new Hadoop cluster for MapR and enable checkbox to use the MapR client.
3. Test the connection as shown below:

You can ignore the Shim Configuration Verification warning, since the value of `fs.defaultFS` does not exist in `core-site.xml`.

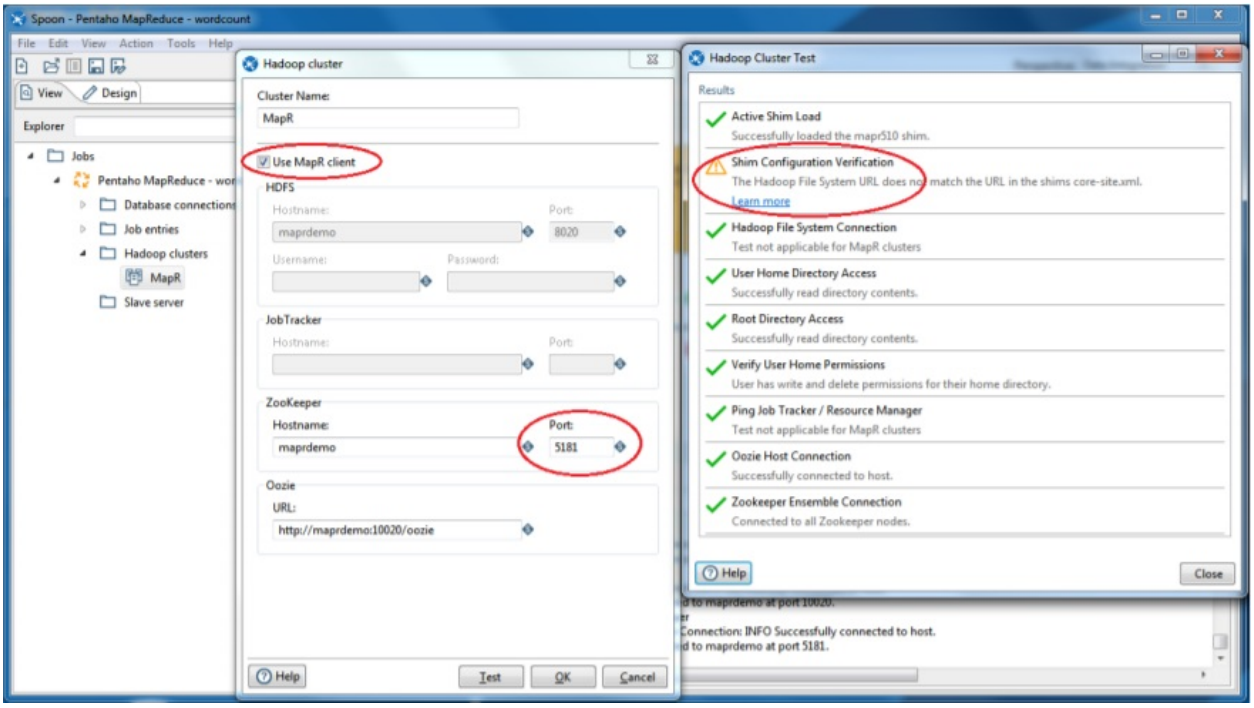

Figure 9: Hadoop Cluster Environment

Running PDI PMR from Samples

Open the sample MapReduce job called `Pentaho MapReduce-wordcount.kjb` found in the `samples\jobs\hadoop` directory.

1. Open the mapping for **Copy Files to HDFS** and change the **Destination** folder as from:

   ```
   hdfs://maprdemo:8020/wordcount/input
   ```

   To:

   ```
   maprfs://maprdemo:8020/wordcount/input
   ```

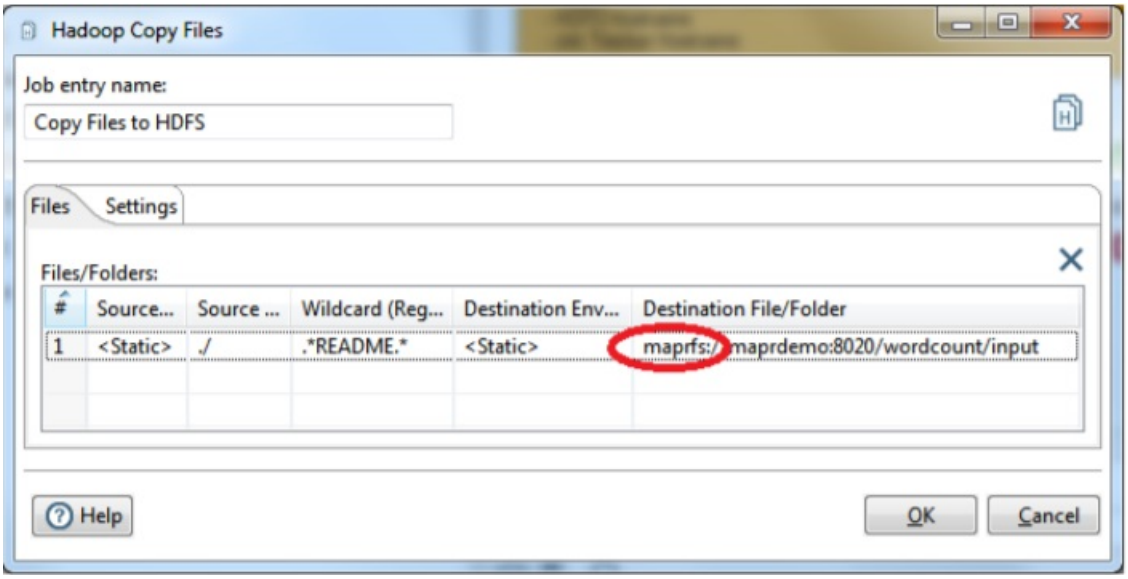

Figure 10: Hadoop Copy Files

2. Run the job and you will see a successful completion as follows:

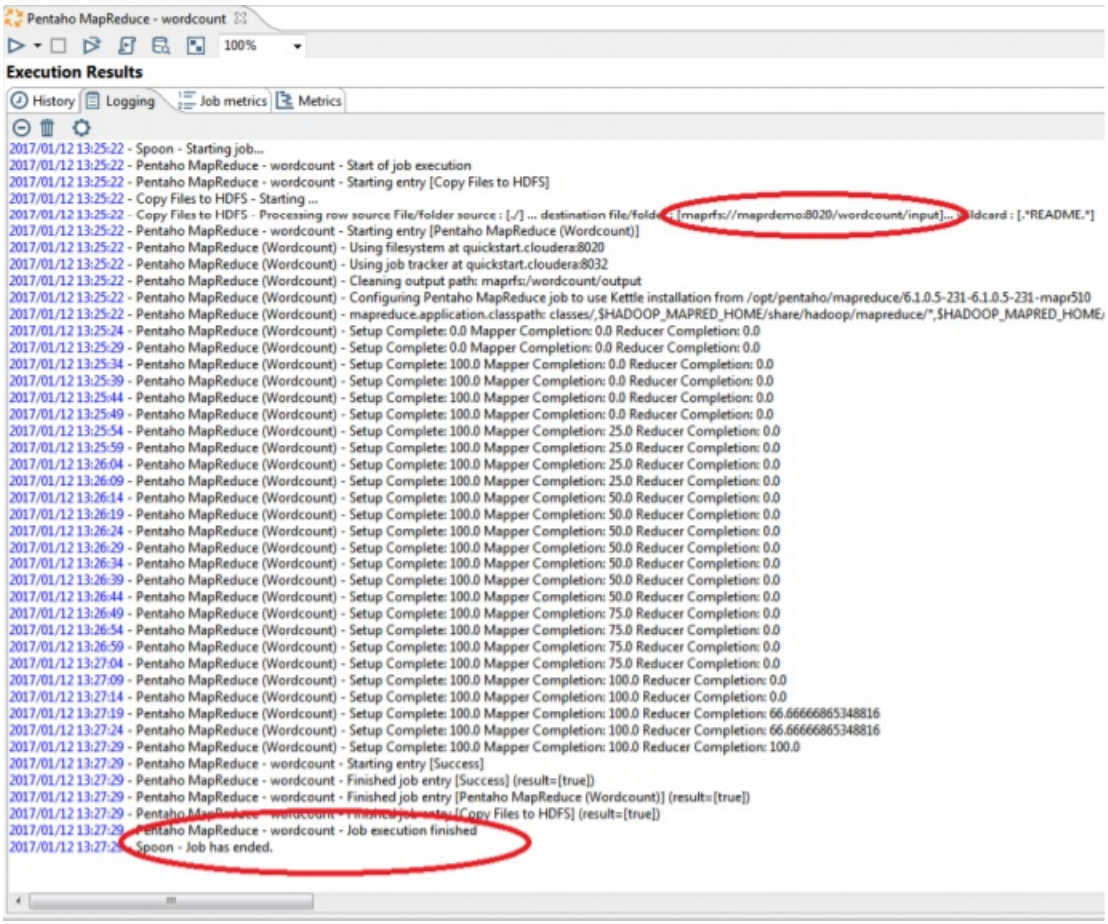

Figure 11: Successful Completion of Job

3. You can confirm that the job has been run successfully by listing the contents of the `hdfs` directory where the files were generated.

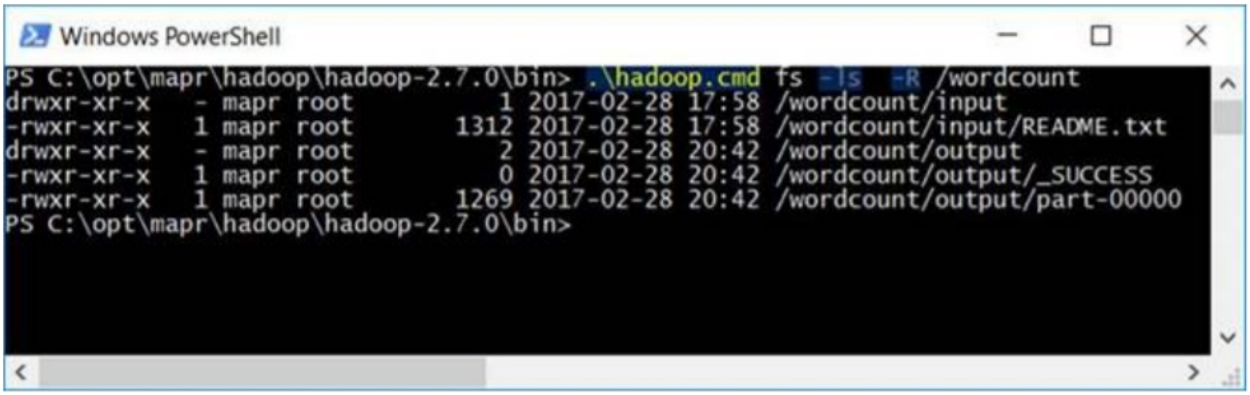

Figure 12: HDFS Directory Confirmation

4. Make sure that the `entry fs.defaultFS` does NOT exist in the `core-site.xml` if you encounter the following error:

```
<property>
```

```
<name>fs.defaultFS</name>
<value>maprfs://maprdemo:8020</value>
<final>true</final>
</property>
```

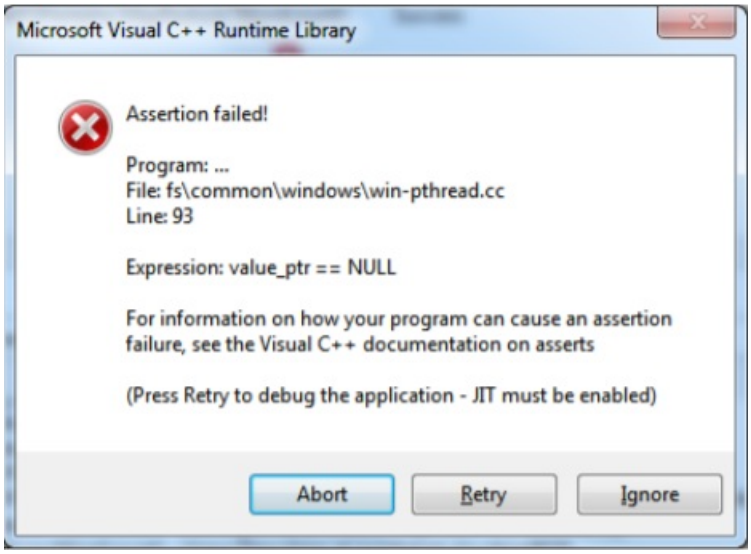

Figure 13: Assertion Failed

Related Information

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

- Index
- MapR Sandbox
- Pentaho Shim for MapR
- Ports Used by MapR
- Pentaho Components Reference

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.

(Sometimes it's good to use a checklist so this is here just in case. Delete if unnecessary.)

Name of the Project:_____

Date of the Review:_____

Name of the Reviewer:_____

Item	Response	Comments
Did you obtain the MapR server information?	YES_____ NO_____	
Did you set up your host environment?	YES_____ NO_____	
Did you download and install the Pentaho Shim for MapR?	YES_____ NO_____	
Did you download the latest MapR client tools from the index?	YES_____ NO_____	
Have you set up the environment variables?	YES_____ NO_____	
Have you configured MapR to connect to HDFS?	YES_____ NO_____	
Did you modify core-site.xml for the MapR client?	YES_____ NO_____	
Did you modify mapred-site.xml for the MapR client?	YES_____ NO_____	
Did you connect to the HDFS using the MapR client?	YES_____ NO_____	
Did you modify `config.properties` in the Pentaho Shim folder?	YES_____ NO_____	
Did you run PDI PMR from samples?	YES_____ NO_____	