Contents

Chapter 1 • Introduction ................................................................. 1
  Audience: SAS and Hadoop Administrators .......................... 1
  What is SAS Data Loader for Hadoop? ................................. 2
  About This Guide ................................................................. 4
  Related Resources ............................................................... 4

Chapter 2 • Start Here: Deployment Scenarios .......................... 5
  How to Use the Scenarios ....................................................... 5
  Scenario 1: First-time Deployment of All SAS Software ............. 6
  Scenario 2: Deployment into an Existing SAS Environment That Uses Hadoop ................................. 7
  Scenario 3: Redeployment After a Hadoop Update .................. 10
  Scenario 4: Switch to a Different Hadoop Distribution ............. 11
  Scenario 5: Upgrade Software .............................................. 13

Chapter 3 • Tasks for the SAS Administrator ............................. 15
  First Steps: SAS Administrators ........................................... 16
  Install SAS Data Loader for Hadoop .................................... 19
  Verify Environment Variables Related to Hadoop .................. 25
  Verify Hadoop Settings ....................................................... 26
  Prepare the SAS Quality Knowledge Base for Deployment ...... 29
  Post-Installation Tasks: SAS Configuration ......................... 30
  Post-Installation Tasks: SAS Data Loader for Hadoop .......... 40
  Define Data Sources .......................................................... 44
  Update SAS Data Loader to Match a New or Updated Hadoop Distribution ........................................ 56
  Update SAS Data Loader Dependencies ............................... 65
  Configure Impala ............................................................... 67

Chapter 4 • Tasks for the Hadoop Administrator ....................... 71
  Hadoop Administrator Checklist ........................................... 72
  Kerberos Administrator Checklist ...................................... 77
  Deploy the SAS Embedded Process ...................................... 79
  Post-Installation Tasks for the SAS Embedded Process ........... 93
  Deploy the SAS Data Loader for Hadoop Spark Engine .......... 98
  Use the Hadoop Tracer Script to Collect Hadoop JAR and Configuration Files .................................... 108
  Deploy the SAS Quality Knowledge Base ......................... 110
  Remove SAS Software from the Hadoop Cluster ................. 119
  Deploy the SAS Data Loader for Hadoop Spark Engine Using
    SAS Deployment Manager ................................................. 125

Recommended Reading ............................................................ 129
Chapter 1
Introduction

Audience: SAS and Hadoop Administrators

The audience for this guide includes both the SAS administrator and the Hadoop administrator. Because SAS Data Loader for Hadoop requires information about the Hadoop cluster and the installation of SAS software on the Hadoop cluster, the SAS administrator must work with a Hadoop administrator. Also, the SAS administrator must work with a Kerberos administrator, if the Hadoop cluster is secured using Kerberos.

SAS Administrator

The SAS administrator must be able to install, configure, administer, and use SAS Intelligence Platform. The general instructions for managing the platform can be found on the SAS Intelligence Platform documentation page: http://support.sas.com/documentation/ onlinedoc/intellplatform/. This guide describes some additional tasks that are required when the platform is installed as part of the SAS Data Loader for Hadoop offering.

The SAS administrator should have the skills to perform the following types of installation, configuration, and administration tasks:

• Use SAS Download Manager to download SAS Software Depot to each machine where the installation will be performed. A thorough understanding of the target network configuration is critical.

• Use SAS Deployment Wizard to install and configure SAS Intelligence Platform and SAS Data Loader for Hadoop.

• Work with the Hadoop administrator to deploy SAS software on the Hadoop cluster using SAS Deployment Manager.

• Use SAS Management Console to maintain the metadata for servers, users, and other global resources, including authorization and credential management permissions.
The SAS Data Loader for Hadoop installation requires information about the Hadoop cluster. A Hadoop administrator must be able to provide information about the Hadoop cluster to the SAS administrator, or be involved more directly with the installation. Also, a Kerberos administrator must perform configuration to support SAS Data Loader for Hadoop specific environments.

Depending on the environment of your Hadoop cluster, information that is required for installation and configuration can include the following:

- Cluster manager credentials
- Location of Hadoop JAR and configuration files

What is SAS Data Loader for Hadoop?

SAS Data Loader for Hadoop is a software offering that makes it easier to move, cleanse, and analyze data in Hadoop. It consists of a web application, elements of the SAS 9.4 Intelligence Platform, and SAS software on the Hadoop cluster.

The SAS Data Loader for Hadoop web application provides an interactive interface that guides you through the process of creating directives. You then run the directives on a SAS Workspace Server. The SAS Workspace Server executes generated code, sends code to Hadoop, and receives responses from Hadoop. During execution, directives provide access to generated code, log information, error messages, and results as they become available. You can save directives, update them, and execute them as needed.
Figure 1.1  SAS Data Loader for Hadoop System Diagram

Multiple users can access the SAS Data Loader for Hadoop web application. The web application uses the SAS Web Infrastructure Platform to connect to SAS servers, to a Hadoop cluster, and to network database management servers. The SAS Metadata Server manages access to data sources and software capabilities for individuals and groups. SAS Management Console provides an administrative interface that configures the web application and manages SAS servers.

SAS software is also deployed to each node in the Hadoop cluster. SAS software on the cluster includes the following components:

- SAS Quality Knowledge Base, which supports data cleansing capabilities in Hadoop
- SAS Embedded Process software, which executes directives in coordination with Hadoop technologies such as Hive, MapReduce, Spark, and Impala
- SAS Data Quality Accelerator, which runs data quality algorithms on Hadoop data
- SAS Data Loader for Hadoop Spark Engine, which executes data integration and data quality tasks in Apache Spark
About This Guide

This guide provides steps for installing and configuring SAS Data Loader for Hadoop. To deploy successfully, start with the Deployment Scenarios on page 5.

Related Resources

- Additional guides are available from the SAS Data Loader for Hadoop documentation page. Business analysts and data stewards can refer to the SAS Data Loader for Hadoop: User's Guide for information about configuring and using directives.

- If you need additional help with deploying or administering the SAS Data Loader for Hadoop, the support community is a great place to find answers.

  SAS Data Management

  Join the community to ask questions and receive expert online support.
Chapter 2

Start Here: Deployment Scenarios

How to Use the Scenarios

The following scenarios represent typical approaches for installing and configuring SAS Data Loader for Hadoop, and related SAS software that is installed on the Hadoop cluster. Each scenario provides an overview of the deployment steps to be performed, by specific IT roles. Here are some tips to help you successfully complete a deployment scenario:

• Print out the scenario to be performed, and refer back to it as you perform the steps.
• Perform the steps in the order shown.
• Each step is associated with one or more of the following IT roles:
  • A SAS administrator, who has the appropriate rights and privileges to install the SAS Intelligence Platform software, including SAS Data Loader for Hadoop.
  • A Hadoop administrator, who has the appropriate rights and privileges to deploy software on a Hadoop cluster. Also, the Hadoop administrator can run a script on the Hadoop NameNode to collect required configuration and JAR files that are required by SAS Data Loader for Hadoop.
  • A Kerberos administrator, who is responsible for managing Kerberos security for the Hadoop cluster that SAS Data Loader for Hadoop uses. Although Kerberos is not required, it is included in the scenarios to help customers who do secure their Hadoop cluster with Kerberos.
Scenario 1: First-time Deployment of All SAS Software

**TIP** Print out this scenario, and schedule a time to go over the steps with the Hadoop administrator. Perform the steps in the order shown.

In this scenario, all SAS software, including SAS Data Loader for Hadoop, is deployed into a new environment.

- SAS Data Loader for Hadoop and SAS Intelligence Platform software is deployed on a single-tier or multi-tier environment.
- A first-time deployment of SAS software on the Hadoop cluster.

**Table 2.1 Steps for a First-time Deployment into a New Environment**

<table>
<thead>
<tr>
<th>PHASE 1: PLANNING</th>
<th>Who Performs It</th>
</tr>
</thead>
<tbody>
<tr>
<td>STEP 1.1:</td>
<td>Create the SAS Software Depot and share ZIP files with the Hadoop administrator.</td>
</tr>
<tr>
<td></td>
<td>Perform first steps on page 16</td>
</tr>
<tr>
<td>STEP 1.2:</td>
<td>Review the Hadoop administrator and Kerberos administrator checklists, as needed.</td>
</tr>
<tr>
<td></td>
<td>Review Hadoop checklist on page 72</td>
</tr>
<tr>
<td></td>
<td>Review Kerberos checklist on page 77</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PHASE 2: GET YOUR HADOOP CLUSTER READY</th>
<th>Who Performs It</th>
</tr>
</thead>
<tbody>
<tr>
<td>STEP 2.1:</td>
<td>Deploy the SAS Embedded Process on the Hadoop cluster.</td>
</tr>
<tr>
<td></td>
<td>Deploy the SAS Embedded Process on page 79</td>
</tr>
<tr>
<td></td>
<td>Perform post-installation configuration for the SAS Embedded Process.</td>
</tr>
<tr>
<td></td>
<td>Configure the SAS Embedded Process on page 93</td>
</tr>
<tr>
<td>STEP 2.2:</td>
<td>Deploy SAS Data Loader for Hadoop Spark Engine on the Hadoop cluster.</td>
</tr>
<tr>
<td></td>
<td>Deploy SAS Data Loader for Hadoop Spark Engine on page 98</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PHASE 3: INSTALL SAS DATA LOADER FOR HADOOP</th>
<th>Who Performs It</th>
</tr>
</thead>
<tbody>
<tr>
<td>STEP 3.1:</td>
<td>Collect the required Hadoop JAR and configuration files.</td>
</tr>
<tr>
<td></td>
<td>Collect the Hadoop client files on page 108</td>
</tr>
<tr>
<td><strong>Note:</strong> Alternatively, the SAS administrator can collect these files by using the SAS Deployment Wizard in the next step.</td>
<td></td>
</tr>
</tbody>
</table>
Scenario 2: Deployment into an Existing SAS Environment That Uses Hadoop

**STEP 3.2:** Run the SAS Deployment Wizard to install SAS Data Loader for Hadoop.

*Install SAS Data Loader for Hadoop on page 19*

**STEP 3.3:** Verify that the paths set for the SAS_HADOOP_JAR_PATH and SAS_HADOOP_CONFIG_PATH environment variables are correct.

*Verify paths for environment variables on page 25*

**STEP 3.4:** Verify Hadoop settings in the inventory.json and ep-config.xml files.

*Verify Hadoop settings on page 26*

► **PHASE 4: DEPLOY YOUR SAS QUALITY KNOWLEDGE BASE (QKB)**

<table>
<thead>
<tr>
<th><strong>STEP 4.1:</strong></th>
<th>Locate the QKB and share it with the Hadoop administrator.</th>
<th>Who Performs It</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prepare the QKB on page 29</td>
<td>SAS administrator</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>STEP 4.2:</strong></th>
<th>Deploy the QKB on the Hadoop cluster.</th>
<th>Hadoop administrator with help from the SAS administrator</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Deploy the QKB on page 110</td>
<td></td>
</tr>
</tbody>
</table>

► **PHASE 5: FINAL STEPS**

<table>
<thead>
<tr>
<th><strong>STEP 5.1:</strong></th>
<th>Update SAS metadata and configuration files as appropriate for your Hadoop distribution.</th>
<th>Who Performs It</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Update metadata and more on page 30</td>
<td>SAS administrator</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>STEP 5.2:</strong></th>
<th>Validate SAS Data Loader for Hadoop and activate the QKB features.</th>
<th>SAS administrator</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Validate and more on page 40</td>
<td></td>
</tr>
</tbody>
</table>

---

**Scenario 2: Deployment into an Existing SAS Environment That Uses Hadoop**

**Tip**  
Print out this scenario, and schedule a time to go over the steps with the Hadoop administrator. Perform the steps in the order shown.

In this scenario, SAS Data Loader for Hadoop is deployed into an existing SAS environment that is already connected to a supported Hadoop cluster.

- An example environment might be where SAS Data Integration Studio uses SAS/ACCESS for Hadoop to read and write data in Hadoop.
In this scenario, SAS Embedded Process and SAS Data Loader for Hadoop Spark Engine software have not yet been installed on the Hadoop cluster.

### Table 2.2  Steps for Deploying into an Existing SAS Environment That Uses a Hadoop Cluster

**► PHASE 1: PLANNING**

<table>
<thead>
<tr>
<th>STEP 1.1:</th>
<th>Create the SAS Software Depot for SAS Data Loader for Hadoop and share ZIP files with the Hadoop administrator.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who Performs It</td>
<td>SAS administrator</td>
</tr>
<tr>
<td><strong>Perform first steps on page 16</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>STEP 1.2:</th>
<th>Review the Hadoop administrator and Kerberos administrator checklists, as needed.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who Performs It</td>
<td>Hadoop administrator, Kerberos administrator</td>
</tr>
<tr>
<td><strong>Review Hadoop checklist on page 72</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Review Kerberos checklist on page 77</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>STEP 1.3:</th>
<th>Review the paths set for the SAS_HADOOP_JAR_PATH and SAS_HADOOP_CONFIG_PATH environment variables.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who Performs It</td>
<td>SAS Administrator</td>
</tr>
<tr>
<td><strong>Review paths for environment variables on page 25</strong></td>
<td></td>
</tr>
</tbody>
</table>

**► PHASE 2: GET YOUR HADOOP CLUSTER READY**

<table>
<thead>
<tr>
<th>STEP 2.1:</th>
<th>Deploy the SAS Embedded Process on the Hadoop cluster.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who Performs It</td>
<td>Hadoop administrator</td>
</tr>
<tr>
<td><strong>Deploy the SAS Embedded Process on page 79</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Perform post-installation configuration for the SAS Embedded Process.</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Configure the SAS Embedded Process on page 93</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>STEP 2.2:</th>
<th>Deploy SAS Data Loader for Hadoop Spark Engine on the Hadoop cluster.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who Performs It</td>
<td>Hadoop administrator</td>
</tr>
<tr>
<td><strong>Deploy SAS Data Loader for Hadoop Spark Engine on page 98</strong></td>
<td></td>
</tr>
</tbody>
</table>

**► PHASE 3: UPDATE EXISTING SOFTWARE AND INSTALL SAS DATA LOADER FOR HADOOP**

<table>
<thead>
<tr>
<th>STEP 3.1:</th>
<th>Collect the required Hadoop JAR and configuration files.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who Performs It</td>
<td>Hadoop administrator</td>
</tr>
<tr>
<td><strong>Collect the Hadoop client files on page 108</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Note:</strong> Alternatively, the SAS administrator can collect these files by using the SAS Deployment Wizard in the next step.</td>
<td></td>
</tr>
</tbody>
</table>
### Scenario 2: Deployment into an Existing SAS Environment That Uses Hadoop

<table>
<thead>
<tr>
<th>Step 3.2: Run the SAS Deployment Wizard to update the existing SAS software and to install SAS Data Loader for Hadoop.</th>
<th>SAS administrator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Note: When you update existing software, during the installation the SAS Deployment Manager starts displaying panels related to SAS Data Loader for Hadoop.</td>
<td></td>
</tr>
<tr>
<td><strong>Install SAS Data Loader for Hadoop on page 19</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step 3.3: Verify that the paths set for the SAS_HADOOP_JAR_PATH and SAS_HADOOP_CONFIG_PATH environment variables are correct.</th>
<th>SAS administrator</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Verify paths for environment variables on page 25</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step 3.4: Verify Hadoop settings in the inventory.json and ep-config.xml files.</th>
<th>SAS administrator with help from the Hadoop administrator</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Verify Hadoop settings on page 26</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>PHASE 4: DEPLOY YOUR SAS QUALITY KNOWLEDGE BASE (QKB)</strong></th>
<th>Who Performs It</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>STEP 4.1: Locate the QKB and share it with the Hadoop administration.</strong></td>
<td>SAS administrator</td>
</tr>
<tr>
<td><strong>Prepare the QKB on page 29</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>STEP 4.2: Deploy the QKB on the Hadoop cluster.</strong></th>
<th>Hadoop administrator with help from the SAS administrator</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Deploy the QKB on page 110</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>PHASE 5: FINAL STEPS</strong></th>
<th>Who Performs It</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>STEP 5.1: Update SAS metadata and configuration files as appropriate for your Hadoop distribution.</strong></td>
<td>SAS administrator</td>
</tr>
<tr>
<td><strong>Update metadata and more on page 30</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>STEP 5.2: Use SAS Management Console to verify and update metadata for SAS Data Loader Dependencies.</strong></th>
<th>SAS administrator</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Update metadata for SAS Data Loader Dependencies on page 65</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>STEP 5.3: Validate SAS Data Loader for Hadoop and activate the QKB features.</strong></th>
<th>SAS administrator</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Validate and more on page 40</strong></td>
<td></td>
</tr>
</tbody>
</table>
Scenario 3: Redeployment After a Hadoop Update

You must redeploy the SAS Data Loader for Hadoop web application for the following reasons:

- You upgrade your Hadoop distribution.
- You install a new Hadoop parcel, package, service, or component on an existing cluster.
- You make any configuration changes to the Hadoop services or components

This scenario assumes the following:

- SAS Data Loader for Hadoop has already been deployed successfully.
- SAS Embedded Process, SAS Data Loader for Hadoop Spark Engine, and a SAS Quality Knowledge Base are already deployed on the Hadoop cluster.

Table 2.3 Steps for Redeploying SAS Data Loader for Hadoop After a Hadoop Update

<table>
<thead>
<tr>
<th>▶ PERFORM THESE STEPS</th>
<th>Who Performs It</th>
</tr>
</thead>
<tbody>
<tr>
<td>STEP 1:</td>
<td></td>
</tr>
<tr>
<td>Apply any hot fixes for your SAS software.</td>
<td>SAS administrator</td>
</tr>
<tr>
<td>Technical Support Hot Fixes site</td>
<td></td>
</tr>
<tr>
<td>STEP 2:</td>
<td></td>
</tr>
<tr>
<td>Review the paths set for the SAS_HADOOP_JAR_PATH and SAS_HADOOP_CONFIG_PATH environment variables are correct.</td>
<td>SAS administrator</td>
</tr>
<tr>
<td>Review paths for environment variables on page 25</td>
<td></td>
</tr>
<tr>
<td>STEP 3:</td>
<td></td>
</tr>
<tr>
<td>Review Hadoop settings in the inventory.json and epconfig.xml files.</td>
<td>SAS administrator</td>
</tr>
<tr>
<td>Review Hadoop settings on page 26</td>
<td></td>
</tr>
<tr>
<td>Note: Review these settings to confirm later that the Hadoop distribution version and related settings are updated.</td>
<td></td>
</tr>
<tr>
<td>STEP 4:</td>
<td></td>
</tr>
<tr>
<td>Collect new Hadoop JAR and configuration (client) files, update SAS metadata for the new Hadoop client files, and redeploy the web application.</td>
<td>SAS Administrator</td>
</tr>
<tr>
<td>Update and deploy on page 56</td>
<td></td>
</tr>
<tr>
<td>Note: Information about the Hadoop cluster is required to complete this step. Also, if you have deployed the SAS server tier and the SAS middle tier on different machines, you must repeat some SAS Deployment Wizard tasks on each machine.</td>
<td></td>
</tr>
</tbody>
</table>
**Scenario 4: Switch to a Different Hadoop Distribution**

*TIP*  Print out this scenario, and schedule a time to go over the steps with the Hadoop administrator. Perform the steps in the order shown.

In this scenario, SAS Data Loader for Hadoop is configured to use a different Hadoop distribution. The following assumptions apply to this scenario:

- SAS Data Loader for Hadoop has already been deployed successfully.
- You want to switch to a different, supported Hadoop distribution. For example, you want to switch from Cloudera to Hortonworks.

<table>
<thead>
<tr>
<th>Table 2.4  Steps for Using a Different Hadoop Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PHASE 1: PLANNING</strong></td>
</tr>
<tr>
<td>STEP 1.1: Apply any hot fixes for your SAS software.</td>
</tr>
<tr>
<td>Technical Support Hot Fixes site</td>
</tr>
<tr>
<td>STEP 1.2: Review the Hadoop administrator and Kerberos administrator checklists, as needed.</td>
</tr>
<tr>
<td>Review Hadoop checklist on page 72</td>
</tr>
<tr>
<td>Review Kerberos checklist on page 77</td>
</tr>
<tr>
<td><strong>PHASE 2: GET YOUR HADOOP CLUSTER READY</strong></td>
</tr>
<tr>
<td>STEP 2.1: If necessary, remove SAS Embedded Process, SAS Data Loader for Hadoop Spark Engine, and the QKB from the target Hadoop cluster.</td>
</tr>
<tr>
<td>Remove software on page 119</td>
</tr>
<tr>
<td>Note: You can also use the same steps to remove the SAS software from the Hadoop cluster that was previously used.</td>
</tr>
</tbody>
</table>
### PHASE 1: DEPLOY SAS EMBEDDED PROCESS FOR HADOOP

**STEP 2.2:** Deploy the SAS Embedded Process on the Hadoop cluster.

- **Deploy the SAS Embedded Process on page 79**
- **Perform post-installation configuration for the SAS Embedded Process.**

**STEP 2.3:** Deploy SAS Data Loader for Hadoop Spark Engine on the Hadoop cluster.

- **Deploy SAS Data Loader for Hadoop Spark Engine on page 98**

<table>
<thead>
<tr>
<th>PHASE 3: CONFIGURE SAS DATA LOADER FOR HADOOP</th>
<th>Who Performs It</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>STEP 3.1:</strong> Review the paths set for the SAS_HADOOP_JAR_PATH and SAS_HADOOP_CONFIG_PATH environment variables.</td>
<td>SAS administrator</td>
</tr>
<tr>
<td><strong>Review paths for environment variables on page 25</strong></td>
<td></td>
</tr>
<tr>
<td><strong>STEP 3.2:</strong> Review Hadoop settings in the inventory.json and ep-config.xml files.</td>
<td>SAS administrator with help from the Hadoop administrator</td>
</tr>
<tr>
<td><strong>Review Hadoop settings on page 26</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Note:</strong> Review these settings to confirm later that the Hadoop distribution version and related settings are updated.</td>
<td></td>
</tr>
<tr>
<td><strong>STEP 3.3:</strong> Collect new Hadoop JAR and configuration (client) files, update SAS metadata for the new Hadoop client files, and redeploy the web application.</td>
<td>SAS administrator with help from the Hadoop administrator</td>
</tr>
<tr>
<td><strong>Update and deploy on page 56</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Note:</strong> Information about the Hadoop cluster is required to complete this step. Also, if you have deployed the SAS server tier and the SAS middle tier on different machines, you must repeat some SAS Deployment Wizard tasks on each machine.</td>
<td></td>
</tr>
<tr>
<td><strong>STEP 3.4:</strong> Verify that the paths set for the SAS_HADOOP_JAR_PATH and SAS_HADOOP_CONFIG_PATH environment variables are correct.</td>
<td>SAS administrator</td>
</tr>
<tr>
<td><strong>Verify paths for environment variables on page 25</strong></td>
<td></td>
</tr>
<tr>
<td><strong>STEP 3.5:</strong> Verify Hadoop settings in the inventory.json and ep-config.xml files.</td>
<td>SAS administrator with help from the Hadoop administrator</td>
</tr>
<tr>
<td><strong>Verify Hadoop settings on page 26</strong></td>
<td></td>
</tr>
</tbody>
</table>

### PHASE 4: DEPLOY YOUR SAS QUALITY KNOWLEDGE BASE (QKB)

<table>
<thead>
<tr>
<th>Who Performs It</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAS administrator with help from the Hadoop administrator</td>
</tr>
</tbody>
</table>
### Scenario 5: Upgrade Software

#### TIP
Print out this scenario, and schedule a time to go over the steps with the Hadoop administrator. Perform the steps in the order shown.

In this scenario, new versions of SAS Embedded Process and SAS Quality Knowledge Base (QKB) are configured for use with SAS Data Loader for Hadoop.

#### Table 2.5  Steps for Upgrading Software

<table>
<thead>
<tr>
<th>PERFORM THESE STEPS</th>
<th>Who Performs It</th>
</tr>
</thead>
</table>
| **STEP 1:** Remove or deactivate SAS Embedded Process.  
“Remove SAS Software from the Hadoop Cluster” on page 119 | Hadoop administrator |
| **STEP 2:** Provide files to the Hadoop administrator.  
Provide files for installing SAS Embedded Process on page 17 | SAS administrator |
| **STEP 3:** Deploy the SAS Embedded Process on the Hadoop cluster.  
Deploy the SAS Embedded Process on page 79 | Hadoop administrator |

<table>
<thead>
<tr>
<th>PHASE 5: FINAL STEPS</th>
<th>Who Performs It</th>
</tr>
</thead>
</table>
| **STEP 5.1:** Perform required post-installation tasks.  
Perform required updates on page 30 | SAS administrator |
| **STEP 5.2:** Verify that Impala is registered in SAS metadata.  
Verify that Impala is registered on page 67  
*Note:* Impala is optional. If Impala was not installed with the previous Hadoop distribution, but it is installed on the new distribution, Impala is not automatically registered. | SAS administrator |
| **STEP 5.3:** Validate SAS Data Loader for Hadoop, and activate the QKB features.  
Validate and activate the QKB on page 40 | SAS administrator |

**STEP 4.1:** Locate the QKB and share it with the Hadoop administration.  
Prepare the QKB on page 29

**STEP 4.2:** Deploy the QKB on the Hadoop cluster.  
Deploy the QKB on page 110

**STEP 5.1:** Perform required post-installation tasks.  
Perform required updates on page 30

**STEP 5.2:** Verify that Impala is registered in SAS metadata.  
Verify that Impala is registered on page 67  
*Note:* Impala is optional. If Impala was not installed with the previous Hadoop distribution, but it is installed on the new distribution, Impala is not automatically registered.

**STEP 5.3:** Validate SAS Data Loader for Hadoop, and activate the QKB features.  
Validate and activate the QKB on page 40
| STEP 4: | Locate the new QKB in the installation folder that was specified in the SAS Deployment Wizard. | SAS administrator |
|        | Prepare the QKB on page 29                                                                 |                |
| STEP 5: | Deploy the QKB on the Hadoop cluster.                                                      | Hadoop administrator with help from the SAS administrator |
|        | “Deploy the SAS Quality Knowledge Base” on page 110                                         |                |
Chapter 3
Tasks for the SAS Administrator

First Steps: SAS Administrators .................................................. 16
  Plan the Deployment with the Hadoop Administrator .................. 16
  Download the SAS Software Depot ........................................ 17
  Provide Files to the Hadoop Administrator ............................. 17
  Provide Information to the Kerberos Administrator ................... 18
  Return to Your Scenario for the Next Step ............................. 19
Install SAS Data Loader for Hadoop ......................................... 19
  Before You Begin ................................................................... 19
  Run the SAS Deployment Wizard .......................................... 21
  Return to Your Scenario for the Next Step ............................. 24
Verify Environment Variables Related to Hadoop ....................... 25
  Before You Begin ................................................................... 25
  Verify SAS Environment Variables for Hadoop ....................... 25
  Return to Your Scenario for the Next Step ............................. 26
Verify Hadoop Settings ............................................................... 26
  Before You Begin ................................................................... 26
  Verify Values in the inventory.json File ................................. 26
  Verify the Hadoop Version in the ep-config.xml File ............... 28
  Return to Your Scenario for the Next Step ............................. 29
Prepare the SAS Quality Knowledge Base for Deployment .......... 29
  About Deploying the QKB ..................................................... 29
  Locate the QKB ................................................................. 30
  Return to Your Scenario for the Next Step ............................. 30
Post-Installation Tasks: SAS Configuration .............................. 30
  Overview ............................................................................. 30
  Other SAS Documentation You Might Need for These Tasks ........ 31
  Update SAS Configuration for Kerberos ................................. 32
  Add at Least One SAS User Account to the Data Loader: Administration Role .................................. 33
  SAS Workspace Server Updates for International Data ............. 34
  SAS Web Application Server Updates for International Data ...... 34
  Updates Required if the SAS Embedded Process was Deployed Using a Script .................................. 35
  Verify the Kerberos Setting for SAS Data Integration Studio .... 36
  Specify winutils.exe in the Hadoop Binary Path for SAS Servers . 36
  Update SAS Configuration for MapR ...................................... 37
  Return to Your Scenario for the Next Step ............................. 39
  Backing Up Saved Directives ................................................. 39
Post-Installation Tasks: SAS Data Loader for Hadoop ............... 40
First Steps: SAS Administrators

Plan the Deployment with the Hadoop Administrator

Approaches to Who Deploys the Software

To deploy the software successfully, you must work with the Hadoop administrator to complete one of the scenarios described in Chapter 2, “Start Here: Deployment Scenarios,” on page 5. Schedule a time to meet with the Hadoop administrator to discuss the following items:

- The following SAS software is installed on the Hadoop cluster: the SAS Embedded Process and the SAS Data Loader for Hadoop Spark Engine. The deployment scenarios recommend that Hadoop administrator install this software prior to you deploying SAS Data Loader for Hadoop web application, and the SAS Intelligence Platform software.

- Hadoop JAR and configuration files (Hadoop client files) are required to build the SAS Data Loader for Hadoop web application before it is deployed. Also, information about the Hive, Impala, and Oozie services is required to deploy the web application.

- As a convenience, you can use the SAS Deployment Wizard to collect the Hadoop client files and other required Hadoop cluster information. To do this, you must have the administrator credentials for a Hadoop cluster manager: Cloudera Manager or Ambari.
If the cluster manager credentials are not available to you, the Hadoop administrator must run a Hadoop tracer script to collect the Hadoop client files before you can deploy SAS Data Loader for Hadoop. When you run the SAS Deployment Wizard to install SAS Data Loader for Hadoop, you must manually enter the location of the files. Also, you must manually enter information about the Hive, Impala, and Oozie services. Consult with the Hadoop administrator to get this information.

Note: Impala is optional.

For information about how SAS and Hadoop administrators can use this guide, see “Audience: SAS and Hadoop Administrators” on page 1.

View System Requirements and Supported Hadoop Distributions

View requirements on support.sas.com.

- System requirements are available from the SAS Data Loader for Hadoop documentation page.

- Supported Hadoop distributions are available from the SAS 9.4 Supported Hadoop Distributions page.

Using MapR

Remember the following approaches to deploying the software in an environment that uses MapR:

- If MapR is your Hadoop distribution, the option to enter Hadoop cluster manager credentials during the SAS software deployment is not available.

- Plan your credentials. Prior to deployment, you must add the users to the host machines where you install the MapR components. To run the MapR client setup on the SAS server and middle tier, you need a list of Container Location Databases (CLDB) nodes for the MapR cluster. Refer to MapR documentation for more information about CLDB.

Using JDBC Drivers

JDBC drivers can be used to access DBMSs. This method is optional, but recommended.

If you decide to use JDBC drivers, identify the DBMSs that SAS Data Loader for Hadoop will access. This helps the Hadoop administrator decide which JDBC drivers to install on the Hadoop cluster. After SAS Data Loader for Hadoop has been deployed, you will install the same JDBC drivers on the SAS Web Application Server on the SAS middle tier. For more information about this task, see “Define JDBC Data Sources” on page 44.

Download the SAS Software Depot

Refer to your SAS Software Order Email (SOE) for instructions on how to download and create the SAS Software Depot. A SAS Software Depot is required for installation.


Provide Files to the Hadoop Administrator

The following software is deployed to the Hadoop cluster before SAS Data Loader for Hadoop is installed. Provide the required files to the Hadoop administrator.
SAS Embedded Process
The required files for installing the SAS Embedded Process software on the Hadoop cluster are different for each release. Provide one of the following files to the Hadoop Administrator.

Table 3.1 SAS Embedded Process: Files by Release

<table>
<thead>
<tr>
<th>SAS Release</th>
<th>Provide This File to the Hadoop Administrator</th>
<th>Note:</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAS 9.4M6 release (November 2018)</td>
<td>Provide the following ZIP file, which is a file attachment to the SOE for the SAS Embedded Process:</td>
<td>This release of the SAS Embedded Process is delivered from a SAS Viya repository, so the ZIP file name and the SOE reference SAS Viya. For all instructions to deploy, configure, upgrade, and uninstall this release of the SAS Embedded Process, follow the instructions in the SAS Embedded Process: Deployment Guide.</td>
</tr>
<tr>
<td>SAS 9.4M5 release (September 2017)</td>
<td>Provide the following ZIP file, which is located in the SAS Software Depot:</td>
<td></td>
</tr>
</tbody>
</table>

SAS Data Loader for Hadoop Spark Engine
Provide the following ZIP file to the Hadoop administrator, which is located in the SAS Software Depot:

SAS Data Loader for Hadoop Spark Engine

Provide Information to the Kerberos Administrator
If your Hadoop cluster and SAS environment will be secured with Kerberos, provide the following information to your Kerberos administrator.

- Go to the SAS 9.4 Intelligence Platform Documentation page, and refer to the following guides for information about connecting SAS software to Kerberos environments:
  - SAS 9.4 Intelligence Platform: Middle-Tier Administration Guide
  - SAS 9.4 Intelligence Platform: Security Administration Guide

- To use a Kerberos environment, SAS Data Loader for Hadoop must be configured properly. The following tasks must be performed:
  - SAS Data Loader for Hadoop must be configured for single sign on to the SAS middle tier with Integrated Windows Authentication (IWA). During the installation of SAS Data Loader for Hadoop, be sure to select IWA as the authentication type.
  - Additional SAS configuration is required after SAS Data Loader for Hadoop is installed. For more information, see “Update SAS Configuration for Kerberos” on page 32.
• If your cluster is on a different Kerberos realm or Active Directory domain than the SAS Data Loader for Hadoop middle tier, SAS server tier, or client PC; you must configure the cluster to map the principals correctly.

• You must use a single realm for Kerberos if you plan to use SAS Data Loader for Hadoop with SAS LASR Analytic Server. SAS Data Loader for Hadoop has a Load Data to LASR directive. This directive can be used to copy Hadoop tables to a SAS LASR Analytic Server. SAS LASR Analytic Servers does not support mixed realms in a trust.

See also “Kerberos Administrator Checklist” on page 77.

Return to Your Scenario for the Next Step

Select one of the following links to return to your scenario.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1: First-time Deployment of All SAS Software</td>
<td>6</td>
</tr>
<tr>
<td>Scenario 2: Deployment into an Existing SAS Environment That Uses Hadoop</td>
<td>7</td>
</tr>
<tr>
<td>Scenario 3: Redeployment After a Hadoop Update</td>
<td>10</td>
</tr>
<tr>
<td>Scenario 4: Switch to a Different Hadoop Distribution</td>
<td>11</td>
</tr>
<tr>
<td>Scenario 5: Upgrade Software</td>
<td>13</td>
</tr>
</tbody>
</table>

Install SAS Data Loader for Hadoop

Before You Begin

Verify System Requirements
View requirements on support.sas.com.

• System requirements are available from the SAS Data Loader for Hadoop documentation page.

• Supported Hadoop distributions are available from the SAS 9.4 Supported Hadoop Distributions page.

Obtain Your Deployment Plan and Review the Pre-installation Checklist
Use the deployment plan for the installation, and refer to the pre-installation checklist before installation.

• The deployment plan, a file named plan.xml, is a summary of the software that you install and configure. When you run the SAS Deployment Wizard, you locate the deployment plan to direct the installation of software.

• You can use a custom deployment plan that is created for your site, or you can choose a standard plan.
• SAS recommends that you store your deployment plans in your SAS Software Depot in the `plan_files` directory.

• For more information, see “About Deployment Plans” in the SAS Intelligence Platform: Installation and Configuration Guide.

• Ensure that you have a SAS Quality Knowledge Base in your plan file.

• Review the pre-installation checklist that is in the `plan_files` directory of your SAS Software Depot. Checklists are bundled with deployment plans, topology diagrams, and other documents.

**Review the SAS Intelligence Platform Documentation**

A SAS administrator installs all of the SAS software shown in the Figure 1.1 on page 3. Most products are installed as they are for other SAS solutions, and some installation tasks are customized for SAS Data Loader for Hadoop.

In addition to this guide, review the SAS 9.4 Intelligence Platform: Installation and Configuration Guide. The SAS 9.4 Intelligence Platform: Installation and Configuration Guide provides a multi-step approach to deploying SAS software. Follow the instructions for your environment, with particular attention to the following:

• Make sure you that understand and set up the required users, groups, and ports. All SAS Data Loader for Hadoop users must have a user account.

• For multiple machine deployments, follow the recommended installation order rules.

*Note:* SAS Pooled Workspace Servers are not supported.

**Get Information from the Hadoop Administrator**

To deploy SAS Data Loader for Hadoop successfully, consider the following:

• Find out which Hadoop distribution that SAS Data Loader for Hadoop will connect to.

• To complete the installation, you must provide the host name for the Hive service, and the host names for the Impala and Oozie services, if you plan to use them. There are two ways to collect this information:

  • For any supported Hadoop distribution, the Hadoop administrator can provide this information and you can manually enter it during installation.

  • For supported Hadoop distributions that use Cloudera Manager or Ambari, you can retrieve this information automatically using the SAS Deployment Wizard. To retrieve this information during installation, you must know the cluster manager user name and password.

*Note:* Impala is optional.

• Locate the Hadoop JAR and configuration files, and copy them to the machine or machines where you will install the SAS server and SAS middle tier.

*Note:* The Hadoop JAR and configuration files must be collected from the Hadoop cluster. A Hadoop administrator can collect these files by using the Hadoop tracer script, before you deploy SAS Data Loader for Hadoop. An alternative approach is collect the files by using the SAS Deployment Wizard, which requires that you have administrator credentials for the Hadoop cluster.

• During installation, you will select and confirm the location of these files for each server tier.

• SAS Data Loader for Hadoop does not deploy successfully if the correct Hadoop JAR and configuration files are missing.
Run the SAS Deployment Wizard

Note: Temporarily disable encryption, anti-virus, anti-malware, anti-spyware, and firewall software. These types of programs can interfere with the deployment process and block files as they are being copied and extracted.

1. Perform one of the following steps to start the wizard:
   - Windows: Start Windows Explorer, and then navigate to the highest-level directory in your SAS Software Depot. Next, right-click `setup.exe`, and then select **Run as administrator**.
   - UNIX: Navigate to the highest-level directory in your SAS Software Depot, and run `.setup.sh`.

   **Note:**
   - Run as a service user that is specific to your installation. Typically, this is the user `sas`. For more information about the different panels, see SAS Intelligence Platform: Installation and Configuration Guide.
   - Run as a user who has Write permissions to install the SAS software.
   - MapR: SAS recommends that you run the installation as the MapR user, typically `mapr`.

   **Tip** Export the display, if needed.

2. Perform the following steps:
   - Answer prompts to step through the screens. Depending on your SAS software order, the SAS installation data file, and the deployment plan, the SAS Deployment Wizard prompts you to perform a variety of tasks.

     **Tip** Allow SAS Deployment Wizard to create a SAS First User account (sasdemo) in metadata during deployment for validation purposes.

   - During the installation, look for the following panels to configure settings that are related or specific to SAS Data Loader for Hadoop.

     **Note:** If you are updating existing software, during the installation the SAS Deployment Manager starts displaying panels related to SAS Data Loader for Hadoop.

     **Tip** For more information about the different SAS Deployment Wizard panels, see SAS Intelligence Platform: Installation and Configuration Guide. Also, for multi-tiered deployments, the panels that are presented depend on which tier you are installing software.

### Table 3.2 Panels Related or Specific to Installing SAS Data Loader for Hadoop

<table>
<thead>
<tr>
<th>Panel</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Select Regional</td>
<td>SAS recommends that you make sure the <strong>Configure as a Unicode</strong></td>
</tr>
<tr>
<td>Settings</td>
<td>server check box is selected.</td>
</tr>
</tbody>
</table>
Select Authentication Type

Make sure you to select the **Integrated Windows Authentication** check box.

- To use Hadoop with Kerberos enabled, SAS Data Loader for Hadoop must be configured for single sign on to the SAS middle tier with IWA.
- Selecting Integrated Windows (IWA) authentication configures SAS Workspace Servers running on Windows or UNIX to use IWA.
- Single sign-on allows a user’s identity, obtained from authentication to the user’s client, to securely pass from the client to processes such as the SAS Metadata Server and the SAS Workspace Server.

SAS Quality Knowledge Base

Two panels are presented. Enter an installation folder, and a license file.

*Tip:* Remember the folder name. After installation, you must provide the folder name to the Hadoop administrator to complete the installation of your SAS Quality Knowledge Base on the Hadoop cluster.

SAS Data Loader Mid-Tier Database Information

Enter the administrator user and password.

Select Hadoop Distribution

Select the Hadoop distribution that SAS Data Loader for Hadoop will connect to.

Use Cluster Manager

Do one of the following:

- Select **None** to manually enter information about the Hadoop cluster.

- Select either **Cloudera Manager** or **Ambari** if you want to use a cluster manager to gather required information from the Hadoop cluster for you. Information about Hive, Impala, and Oozie services can be retrieved using the cluster manager.

  *Note:* The Cloudera Manager option can be selected for supported versions of Cloudera CDH. The Ambari option can be selected for supported versions of Hortonworks HDP and IBM BigInsights.

If you select **Cloudera Manager** or **Ambari**, you must complete the next two panels: **Hadoop Cluster Manager Information** and **Hadoop Cluster Manager Credentials**.

*Note:* MapR: If you selected MapR as your Hadoop distribution, this panel might not be displayed.
<table>
<thead>
<tr>
<th>Panel</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hadoop Cluster Manager Information</strong></td>
<td>Enter the following information:</td>
</tr>
<tr>
<td></td>
<td>• Enter the host name and port number for your Hadoop cluster manager.</td>
</tr>
<tr>
<td></td>
<td>For Cloudera, enter the location where Cloudera Manager is running. For Hortonworks, IBM BigInsights, or Pivotal, enter the location where the Ambari server is running. The port number is set to the appropriate default after Cloudera, Hortonworks, IBM BigInsights, or Pivotal is selected above.</td>
</tr>
<tr>
<td></td>
<td>• Indicate whether the cluster manager is enabled with Transport Layer Security (TLS).</td>
</tr>
<tr>
<td></td>
<td>Note: The host name must be a fully qualified domain name. The port number must be valid, and the cluster manager must be listening. The SAS Deployment Manager validates a constructed URL based on the host name and port number and issues an error message if the URL is not accessible.</td>
</tr>
<tr>
<td></td>
<td>Note: All discussion of TLS is also applicable to the predecessor protocol, Secure Sockets Layer (SSL).</td>
</tr>
<tr>
<td><strong>Hadoop Cluster Manager Credentials</strong></td>
<td>Enter the Cloudera Manager or Ambari administrator account name and password.</td>
</tr>
</tbody>
</table>
Enter or confirm the following information:
• Enter the host names for the Hive, Impala, and Oozie services. Impala is an optional service. Note: As needed, contact your Hadoop administrator for this information.
• Specify how to collect the Hadoop JAR and configuration files.
  • Select Manual Collection to manually enter the location of the files.
  • Select SAS Deployment Wizard — Password or SAS Deployment Wizard — Private Key File to collect the files automatically from the Hadoop cluster. To continue to the next panels, you must have administrator credentials to continue, such as the Hadoop cluster SSH credentials. Note: As needed, contact your Hadoop administrator for this information.

Note: If you experience problems connecting to Hadoop and collecting the files, contact SAS Technical Support and include the hadoopgtracer.log file for reference. Look in the following locations for the hadooptracer.log file:
• In the SAS-configuration-directory/HadoopServer/conf/directory on the machine where you ran the SAS Deployment Wizard.
• In either the home directory or the /tmp directory of the UNIX account with SSH.

Enter locations of the Hadoop client configuration files and JAR files. Note: As needed, contact your Hadoop administrator for this information.

Select this option to add two Hadoop cluster environment variables to the SAS configuration file sasv9_usermods.cfg. Tip: Refer to the panel help for more information about this file.

3. Continue through the remaining panels.

Note: If you are performing a multi-tier deployment, run the SAS Deployment Wizard again, as needed, to complete the installation.

Return to Your Scenario for the Next Step
Select one of the following links to return to your scenario.

“Scenario 1: First-time Deployment of All SAS Software” on page 6

“Scenario 2: Deployment into an Existing SAS Environment That Uses Hadoop” on page 7

“Scenario 3: Redeployment After a Hadoop Update” on page 10
Verify Environment Variables Related to Hadoop

Before You Begin

- As a best practice, perform this task after you install or update SAS Data Loader for Hadoop.
- Perform this task to verify the following:
  - That the paths set for the SAS_HADOOP_JAR_PATH and SAS_HADOOP_CONFIG_PATH environment variables are correct.
  - That the path for each environment variable is listed once.

Note: On the SAS server tier, SAS_HADOOP_JAR_PATH sets the path to the Hadoop JAR files, and SAS_HADOOP_CONFIG_PATH sets the path to the Hadoop client configuration files.

Verify SAS Environment Variables for Hadoop

Perform the following steps:

1. Invoke SAS:
   - UNIX:
     1. Open a command shell on the host for the SAS Data Loader Workspace Server.
     2. Run the command:
        
        ```
        /usr/local/sas/config/Lev1/SASApp/WorkspaceServer/
        WorkspaceServer.sh
        ```
   - Windows: Run the following command as administrator on the host for the SAS Data Loader Workspace Server:
     ```
     C:\SAS\Lev1\SASApp\WorkspaceServer\WorkspaceServer.bat
     ```

2. Submit the following code:
   ```
   PROC options; run;
   ```

3. Verify that the paths for SAS_HADOOP_JAR_PATH and SAS_HADOOP_CONFIG_PATH are listed only once in the SAS log.

The following example shows duplicate listings for each environment variable in the SAS log.

```
Duplicate listings can happen if multiple configuration files set these environment variables. You can check by opening `sasv9*.cfg` files in the `SASConfig` and `SASInstall` directories and subdirectories and looking for the statements such as the following.

-SET SAS_HADOOP_CONFIG_PATH "C:\additional-location\conf"

4. Remove any duplicate listings, as needed.

For more information about setting these variables, see the “SAS Environment Variables for Hadoop” appendix in SAS Hadoop Configuration Guide for Base SAS and SAS/ACCESS. This book can be accessed from the SAS 9.4 Support for Hadoop page on support.sas.com.

Return to Your Scenario for the Next Step

Select one of the following links to return to your scenario.

- “Scenario 1: First-time Deployment of All SAS Software” on page 6
- “Scenario 2: Deployment into an Existing SAS Environment That Uses Hadoop” on page 7
- “Scenario 3: Redeployment After a Hadoop Update” on page 10
- “Scenario 4: Switch to a Different Hadoop Distribution” on page 11
- “Scenario 5: Upgrade Software” on page 13

Verify Hadoop Settings

Before You Begin

- Perform the following tasks to confirm that the SAS Data Loader web application and SAS Embedded Process are configured correctly such that they can use your Hadoop distribution.
- As a best practice, perform these tasks after you install or update SAS Data Loader for Hadoop, and after you update your Hadoop distribution.
- To complete the following tasks, information about your Hadoop distribution is required. Consult with your Hadoop administrator.

Verify Values in the `inventory.json` File

The `inventory.json` file contains Hadoop settings for the SAS Data Loader web application.

Perform the following steps to verify the settings:

1. Open the `inventory.json` file to view it:
• UNIX:
  1. Open a command shell on the host for the SAS Data Loader web application.
  2. Use an editor to open the file. A typical path to the file is `SAS-installation-directory/Lev1/Web/Applications/DataLoader/Hadoop/Configs/inventory.json`
• Windows: Open the inventory.json file. A typical path to the file is `SAS-installation-directory\Lev1\Web\Applications\DataLoader\Hadoop\Configs\inventory.json`

2. View the contents of the file. An example of the contents is shown below.

_Note:_ The items in quotation marks differ based on the distribution and features of your Hadoop cluster.

```json
{
  "hadoop_distro": "Distribution Name",
  "hadoop_internal_name": "distribution internal name",
  "hadoop_distro_rpm": "Distribution Name",
  "hadoop_distro_version": "Distribution Name",
  "hadoop_distro_yum": "Distribution Short Name",
  "hive": {
    "port": "10000",
    "hosts": ["hiveserver.MyDomain.com"]
  },
  "impala": {
    "available": "true",
    "port": "21050",
  },
  "spark": {
    "available": "true"
  },
  "oozie": {
    "available": "true",
    "url": "http://oozieserver.MyDomain.com:11000/oozie"
  },
  "hcatalog": {
    "classpath": "/etc/hadoop/conf:/opt/cloudera/parcels/jar5.8.0.p0.42/lib/hadoop"
  },
  "sas_dmarchitect": {
    "available": "true",
    "sas_dm_home": "/usr/local/sas/SASDMPHome"
  }
}
```

3. Verify the contents of the file, including the following:
   • Make sure that the content is valid JSON format.
   • For Spark, make sure that the following values are correct:
     • If the `spark available` value is `false`, and you expect to use Spark for execution, this might indicate that the Hadoop cluster is not configured properly for Hadoop Spark.
     • Make sure that the value for `hcatalog classpath` is a single line, and does not contain a new line or line break.
• Make sure that `sas_dmarchitech` is available and that the installation path is correct. If either is not true, SAS Data Loader for Hadoop Spark Engine might not be installed.

Note: If any of the preceding settings is incorrect, the Oozie job used to run the Spark driver can fail.

• For MapR, it is important that the following values are correct:
  • `hadoop_distro`
  • `hadoop_internal_name settings`

**Verify the Hadoop Version in the ep-config.xml File**

When SAS Embedded Process is installed, the `ep-config.xml` file is created and added to the HDFS `sas/ep/config` directory. The `ep-config.xml` file contains Hadoop settings for SAS Embedded Process.

Perform the following steps to verify that SAS Embedded Process is configured to use the version of your Hadoop distribution.

1. Run the following command to view the file from a host that has access to the Hadoop cluster:
   ```bash
   hdfs dfs -cat /sas/ep/config/ep-config.xml
   ```
2. Search for the `sas.ep.classpath` property, and then confirm the version of the Hadoop distribution. The version is part of the path within the associated value of the `sas.ep.classpath` property.

   In the following example, the version is 2.5.0.0-1245.
   ```xml
   <name>sas.ep.classpath</name>
   <value>$HADOOP_CONF_DIR,/opt/SASEPHome/jars/sasep.jar,
   /usr/hdp/2.5.0.0-1245/hive-hcatalog/libexec/../../share/hcatalog/*,
   /usr/hdp/2.5.0.0-1245/hive/lib/*</value>
   ```
3. Perform one of the following steps:

   • If the version for `sas.ep.classpath` matches your Hadoop distribution, continue to “Return to Your Scenario for the Next Step” on page 29.

   • If the version for `sas.ep.classpath` does not match your Hadoop distribution, run the `sasep-admin.sh` script with the `-genconfig` option.

   Note:

   • The `sasep-admin.sh` script is deployed with the SAS Embedded Process software, and the directory where the script is located differs by the SAS Embedded Process release. For more information, see details about the different SAS Embedded Process releases on page 72 in the Hadoop administrator checklist.

   • If the instructions in this guide were used to deploy the SAS Embedded Process, then see “Run the Script” on page 83.

   • If the instructions in the SAS Embedded Process: Deployment Guide were used to deploy the SAS Embedded Process, then use the instructions in that guide to run the `sasep-admin.sh` script.
Prepare the SAS Quality Knowledge Base for Deployment

About Deploying the QKB

Data cleansing directives in SAS Data Loader for Hadoop reference a SAS Quality Knowledge Base (QKB). To use a QKB, you must deploy it to all DataNodes of the Hadoop cluster.

There are two general steps for deploying a QKB:

1. Locate the QKB that you want to use. To deploy the QKB to the Hadoop cluster, you must know the name of the directory that contains the QKB. For more information, see “Locate the QKB” on page 30.

2. Deploy the QKB to the Hadoop cluster by performing one of the following methods. Typically, a Hadoop administrator deploys the QKB to the Hadoop cluster.
   
   • Deploy a parcel or stack of the QKB by using Cloudera Manager or Ambari, respectively. To create the parcel or stack and make it available to Cloudera Manager or Ambari, you must use the SAS Deployment Manager.

   Note: This method works only if your Hadoop cluster and SAS server tier are based on UNIX.

   • Run a script. This method works for all supported Hadoop distributions and does not require the use of the SAS Deployment Manager. To deploy the QKB using a script, be sure to provide the location of the QKB to the Hadoop administrator who will deploy it.

   Note: If you want to replace a QKB on the Hadoop cluster, you must remove the old one first. For more information, see “Remove SAS Software from the Hadoop Cluster” on page 119.
**Locate the QKB**

There are three ways to locate a QKB for use in SAS Data Loader for Hadoop.

- Use the QKB that was installed with SAS Data Loader for Hadoop. To locate that QKB, navigate to the QKB installation folder that was specified in the SAS Deployment Wizard when you installed SAS Data Loader for Hadoop on page 19. The QKB is typically installed on the SAS Workspace Server or on the same machine that the SAS Software Depot was installed. The administrator who installed your SAS software can identify the machine where the QKB was installed.

  > Note: Most sites use the QKB that was installed with SAS Data Loader for Hadoop.

- Use a customized QKB or some other QKB that you already use with other SAS software in your enterprise.

- If you have a SAS profile, you can get a new or updated QKB from the SAS downloads site. Perform the following steps:
  1. Open the SAS Downloads site.
  2. Select the appropriate QKB.
  3. When prompted, log on to your SAS profile or create a new profile.
  4. Complete downloading and installing a local copy of the QKB.

**Return to Your Scenario for the Next Step**

Select one of the following links to return to your scenario.

- “Scenario 1: First-time Deployment of All SAS Software” on page 6
- “Scenario 2: Deployment into an Existing SAS Environment That Uses Hadoop” on page 7
- “Scenario 3: Redeployment After a Hadoop Update” on page 10
- “Scenario 4: Switch to a Different Hadoop Distribution” on page 11
- “Scenario 5: Upgrade Software” on page 13

---

**Post-Installation Tasks: SAS Configuration**

**Overview**

A SAS administrator should review this section immediately after SAS Data Loader for Hadoop has been installed. Some of these tasks are required to make SAS Data Loader for Hadoop fully operational. Other tasks might be required for your Hadoop distribution and the features that you want to use.
### Table 3.3  Post-Installation Tasks: SAS Configuration

<table>
<thead>
<tr>
<th>Task and Documentation</th>
<th>When Required</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Verify Environment Variables Related to Hadoop” on page 25</td>
<td>This task is not required, but it is a best practice. Perform this task before and after you install or update SAS Data Loader for Hadoop.</td>
</tr>
<tr>
<td>“Verify Hadoop Settings” on page 26</td>
<td>This task is not required, but it is a best practice. Perform this task after you install or update SAS Data Loader for Hadoop.</td>
</tr>
<tr>
<td>“Update SAS Configuration for Kerberos” on page 32</td>
<td>If your cluster is secured with Kerberos, perform these tasks.</td>
</tr>
<tr>
<td>“Add at Least One SAS User Account to the Data Loader: Administration Role” on page 33</td>
<td>If you have not specified at least one Data Loader Administrator in SAS Management Console, perform this task.</td>
</tr>
<tr>
<td>“SAS Workspace Server Updates for International Data” on page 34</td>
<td>Perform these updates on the SAS Workspace Server to reduce errors from international data and from special characters in table and column names.</td>
</tr>
<tr>
<td>“SAS Web Application Server Updates for International Data” on page 34</td>
<td>Perform these updates on the SAS Web Application Server if you plan to submit SAS code in languages other than English.</td>
</tr>
<tr>
<td>“Updates Required if the SAS Embedded Process was Deployed Using a Script” on page 35</td>
<td>Perform these updates if the SAS Embedded Process was deployed using a script.</td>
</tr>
<tr>
<td>“Verify the Kerberos Setting for SAS Data Integration Studio” on page 36</td>
<td>Perform this update if saved directives from SAS Data Loader for Hadoop are included in SAS Data Integration Studio jobs.</td>
</tr>
<tr>
<td>“Specify winutils.exe in the Hadoop Binary Path for SAS Servers” on page 36</td>
<td>If your SAS Workspace Server or SAS Web Application Server is based on Windows and your Hadoop cluster is based on UNIX, perform this task.</td>
</tr>
<tr>
<td>“Update SAS Configuration for MapR” on page 37</td>
<td>If you have a MapR cluster, perform these tasks.</td>
</tr>
<tr>
<td>“Define Data Sources” on page 44</td>
<td>If you want to move data between Hadoop and other systems, perform these tasks.</td>
</tr>
<tr>
<td>“Backing Up Saved Directives” on page 39</td>
<td>Perform this task when you want to back up saved directives. This is a task that you might do after SAS Data Loader for Hadoop has been in operation for a while.</td>
</tr>
</tbody>
</table>

### Other SAS Documentation You Might Need for These Tasks

Most of the books cited in this section are available from the administration documentation page for the SAS Intelligence Platform.
Some tasks in this section require a SAS administrator to use SAS Management Console to add or update metadata. If you are unfamiliar with SAS Management Console, see the documentation page for SAS Management Console on support.sas.com. New users should start with the "Administering SAS Management Console" chapter in SAS 9.4 Intelligence Platform: Desktop Application Administration Guide. For information about managing users and groups, see the SAS 9.4 Management Console: Guide to Users and Permissions.

**Update SAS Configuration for Kerberos**

**Overview**

If your cluster is secured with Kerberos, always perform these tasks.

**Enable Support for Integrated Windows Authentication**

If your cluster is secured with Kerberos, the SAS middle tier must be configured for single sign-on with Integrated Windows Authentication (IWA). This is a standard SAS task. See “Support for Integrated Windows Authentication” in the SAS 9.4 Intelligence Platform: Middle-Tier Administration Guide.

After you enable IWA on the SAS middle tier, you must enable the middle tier to communicate with the Hadoop cluster. To do that, specify a JVM option in a configuration file for each instance of the SAS Web Application Server for SAS Data Loader for Hadoop.

Perform the following steps:

1. Open a command prompt on the SAS Web Application Server for SAS Data Loader for Hadoop.
   
   SAS Data Loader for Hadoop has a dedicated SAS Web Application Server. By default, this server is identified as SASServer15_n in a list of running processes. It is installed in a directory such as SAS-configuration-directory\Lev1\Web \WebAppServer\SASServer15_1. For more information about accessing web application servers, see the “Understanding SAS Web Application Server Configuration” topic in the SAS 9.4 Intelligence Platform: Middle-Tier Administration Guide.

2. Navigate to the folder of configuration files where JVM options can be specified. For more information about specifying JVM options, see the “Understanding SAS Web Application Server Configuration” topic in the SAS 9.4 Intelligence Platform: Middle-Tier Administration Guide.

3. Add the following attribute to the appropriate configuration file:
   
   -Djavax.security.auth.useSubjectCredsOnly=false

4. Save your changes.

5. Repeat for all SAS Web Application Servers for SAS Data Loader for Hadoop (all SASServer15_n servers).

6. Restart the servers that you updated.

   For more information about restarting these servers, see the “Understanding SAS Web Application Server Management” topic in the SAS 9.4 Intelligence Platform: Middle-Tier Administration Guide.
Verify That SAS Data Loader Users Do Not Have a Saved Password for Any Login to DefaultAuth

If your Hadoop cluster is secured with Kerberos, SAS Data Loader for Hadoop must be configured for single sign-on to the SAS middle tier, using Integrated Windows Authentication (IWA). In this environment, the SAS accounts for SAS Data Loader for Hadoop users should not have a saved password of any login to the DefaultAuth authentication domain.

The next figure shows an example of a user (dbtest) with a saved password for a personal login to the DefaultAuth authentication domain. This is not supported on a SAS middle tier configured for single sign-on, using IWA.

Figure 3.1  Personal Login with a Saved Password to the DefaultAuth Authentication Domain

The next figure shows a group (SAS General Servers) with a saved password for a group login to the DefaultAuth authentication domain. SAS Data Loader for Hadoop users should not be members of such a group because if they are, single sign-on does not work properly.

Figure 3.2  Group Login with a Saved Password to the DefaultAuth Authentication Domain

For information about managing users and groups, see the SAS 9.4 Management Console: Guide to Users and Permissions. This book is available from the documentation page for SAS Management Console on support.sas.com.

Add at Least One SAS User Account to the Data Loader: Administration Role

At least one SAS user should be assigned to the Data Loader: Administration role in SAS Management Console. A member of this role has two privileges in the SAS Data Loader that other users do not have:
In the Run Status directive, a Data Loader Administrator can control all jobs that have been submitted by all users.

In the Configuration window, a Data Loader Administrator has access to the QKB Administration panel. This panel enables the administrator to synchronize QKB metadata on your Hadoop cluster with the SAS middle tier. This synchronization enables the main data quality features in SAS Data Loader for Hadoop.

Perform these steps to add an existing SAS user to the Data Loader: Administration role in SAS Management Console:

1. Log on to SAS Management Console as an administrator.
2. From the Plugins tab, select User Manager. User accounts are displayed in the right pane.
3. Right-click the appropriate user account in the right pane and select Properties.
4. In the properties dialog box, click the Groups and Roles tab. Available groups and roles are displayed in the left pane.
5. Select the Data Loader: Administration role in the left pane.
6. Move the selected role to the right pane.
7. Click OK to save your changes. These changes take effect the next time the user logs on to SAS Data Loader for Hadoop.

SAS Workspace Server Updates for International Data

Perform these updates on the SAS Workspace Server to reduce errors from international data and from special characters in table and column names.

Set the following options in the sasv9_usermods.cfg file for the SAS Workspace Server. After you make these changes, restart the server or servers to make the changes go into effect.

- ENCODING UTF8
- VALIDVARNAME ANY
- VALIDMEMNAME EXTEND

The path to the sasv9_usermods.cfg file is similar to the following examples.

Windows: SAS-configuration-directory\Lev1\server-context\sasv9_usermods.cfg
UNIX: SAS-configuration-directory/Lev1/server-context/sasv9_usermods.cfg

The ENCODING option reduces errors that can occur when data in one language is shared with computers that operate in another language. The VALIDVARNAME option reduces errors that can occur when the names for tables and columns include spaces and special characters. The VALIDMEMNAME option increases the valid characters that are allowed for the names of tables and columns.

SAS Web Application Server Updates for International Data

The directive Run a SAS Program provides the primary means of submitting user-written SAS code in SAS Data Loader for Hadoop. The code runs as you submit it, without the code generation step that is used in other directives. The code that you submit generates the same log and error information as in other directives.
If you plan to submit SAS code in a language other than English, update the following configuration files for the SAS Web Application Server for SAS Data Loader for Hadoop. These updates reduce transcode errors in the submitted SAS code and the log. After you make both changes, restart the server to make the changes go into effect.

SAS Data Loader for Hadoop has a dedicated SAS Web Application Server. By default, this server is identified as SASServer15_n in a list of running processes. It is installed in a directory such as SAS-configuration-directory\Lev1\Web\WebAppServer\SASServer15_1. For more information about accessing web application servers, see the “Understanding SAS Web Application Server Configuration” topic in the SAS 9.4 Intelligence Platform: Middle-Tier Administration Guide.

Set the following option in the sasv9_usermods.cfg file for all SAS Web Application Servers for SAS Data Loader for Hadoop.

-ENCODING UTF8

The path to the sasv9_usermods.cfg file is similar to the following examples.

Windows: SAS-configuration-directory\Lev1\Web\WebAppServer\SASServer15_1\sasv9_usermods.cfg
UNIX: SAS-configuration-directory/Lev1/Web/WebAppServer/SASServer15_1/sasv9_usermods.cfg

The ENCODING option reduces errors that can occur when data in one language is shared with computers that operate in another language.

Set the following JVM option in the wrapper.conf file for the SAS Web Application Server:

wrapper.java.additional.51=-Dfile.encoding=utf-8

The path to the wrapper.conf file is similar to the following examples.

Windows: SAS-configuration-directory\Lev1\Web\WebAppServer\SASServer15_1\conf/wrapper.conf
UNIX: SAS-configuration-directory/Lev1/Web/WebAppServer/SASServer15_1/conf/wrapper.conf

Updates Required if the SAS Embedded Process was Deployed Using a Script

Note: If you used the SAS Deployment Manager to install the SAS Embedded Process, these configuration tasks are not necessary. They were completed using the SAS Deployment Manager.

The following additional configuration tasks must be performed:

• The hive-site.xml configuration file must be in the SAS_HADOOP_CONFIG_PATH.

• The following Hive or HCatalog JAR files must be in the SAS_HADOOP_JAR_PATH.

hive-hcatalog-core-*.jar
hive-webhcat-java-client-*.jar
jdo-api*.jar

• If you are using MapR, the following Hive or HCatalog JAR files must be in the SAS_HADOOP_JAR_PATH.
To access Avro file types, the avro-1.7.4.jar file must be added to the SAS_HADOOP_JAR_PATH environment variable.

To access Parquet file types with Cloudera 5.1, the parquet-hadoop-bundle.jar file must be added to the directory defined in the SAS_HADOOP_JAR_PATH environment variable.

If your distribution is running Hive 0.12, the jersey-client-1.9.jar must be added to the SAS_HADOOP_JAR_PATH environment variable.

For more information about the SAS_HADOOP_JAR_PATH and SAS_HADOOP_CONFIG_PATH environment variables, see the SAS Hadoop Configuration Guide for Base SAS and SAS/ACCESS.

Verify the Kerberos Setting for SAS Data Integration Studio

If saved directives from SAS Data Loader for Hadoop are included in SAS Data Integration Studio jobs, verify that the SAS Data Loader is secured with Kerberos field is accurate for your Hadoop cluster. This field enables SAS Data Integration Studio to generate the appropriate authentication code for a job that includes a saved directive.

If your cluster is secured with Kerberos, set this field to True. Otherwise, set this field to False.

Perform the following steps:

1. Log on to SAS Management Console as an administrator.
2. From the Plugins tab, select: Application Management > Configuration Manager > Data Loader Midtier Shared Services.
3. Right-click Data Loader Midtier Shared Services and select Properties.
4. Click the Settings tab.
5. Select SAS Data Loader Dependencies in the left pane.
6. Review the SAS Data Loader is secured with Kerberos field in the right pane.
   
   If your cluster is secured with Kerberos, set this field to True. Otherwise, set this field to False.
7. Click OK to save any changes in SAS Management Console.

The SAS Web Application Server does not have to be restarted in this case. The SAS Data Loader is secured with Kerberos field is used by SAS Data Integration Studio, which does not use the SAS Web Application Server.

Specify winutils.exe in the Hadoop Binary Path for SAS Servers

Perform this task if your SAS Workspace Server or SAS Web Application Server are based on Windows, and your Hadoop cluster is based on UNIX. These servers are not able to connect to the Hadoop cluster unless you specify winutils.exe in their Hadoop
binary path. They generate a JAVA exception that says that the winutils.exe file cannot be found.

You can use the HADOOP_HOME variable to specify the Hadoop binary path for the SAS Workspace Server and the SAS Web Application Server for SAS Data Loader for Hadoop.

Perform these steps:

1. Download the winutils.exe file for your Hadoop distribution.
2. Display a command prompt on a SAS Workspace Server or a SAS Web Application Server for SAS Data Loader for Hadoop.

By default, the SAS Web Application Server for SAS Data Loader for Hadoop is identified as SASServer15_n in a list of running processes. It is installed in a directory such as SAS-configuration-directory\Lev1\Web\WebAppServer\SASServer15_1. For more information about accessing web application servers, see the “Understanding SAS Web Application Server Configuration” topic in the SAS 9.4 Intelligence Platform: Middle-Tier Administration Guide.

3. Create a folder, such as c:\winutils\bin.
4. Copy winutils.exe to the folder you just created.
5. Add a line such as the following to the configuration file for the SAS server: set path HADOOP_HOME c:\winutils
6. Save your changes.
7. Repeat for any other SAS Workspace Server or SAS Web Application Server for SAS Data Loader for Hadoop (other SASServer15_n servers).
8. Restart the servers that you updated.

For more information about restarting web application servers, see the “Understanding SAS Web Application Server Management” topic in the SAS 9.4 Intelligence Platform: Middle-Tier Administration Guide.

Update SAS Configuration for MapR

Overview
If you have a MapR cluster, always perform these tasks.

Add a MapR Login Option to the sasv9_usermods.cfg File
If you have a MapR cluster, you must add a MapR login option to the sasv9_usermods.cfg file for the SAS Application Server for SAS Data Loader for Hadoop.

Perform the following steps:

1. Navigate to the folder that contains the sasv9_usermods.cfg file. Here are the typical paths to the sasv9_usermods.cfg file:
   - UNIX: SAS-configuration-directory/Lev1/server-context/sasv9_usermods.cfg
   - Windows: SAS-configuration-directory\Lev1\server-context\sasv9_usermods.cfg
2. Add a MapR login option to the sasv9_usermods.cfg file. Here is an example of that option:

    -jreoptions (-Djava.security.auth.login.config=SAS-configuration-directory/Lev\n    n/HadoopServer/conf/mapr.login.conf)

**Configure the SAS Web Application Server for MapR**

If you have a MapR cluster, you must perform the following configuration for the SAS Web Application Server for SAS Data Loader for Hadoop:

- update the JAAS configuration file for the SAS Web Application Server for SAS Data Loader for Hadoop
- update the setenv file to set impersonation and to specify the MAPR_HOME location

Perform the following steps to update the JAAS configuration file:

1. Display a command prompt on the SAS Web Application Server for SAS Data Loader for Hadoop.

   By default, this server is identified as SASServer15_n in a list of running processes. It is installed in a directory such as SAS-configuration-directory\Lev1\Web\WebAppServer\SASServer15_1. For more information about accessing web application servers, see the “Understanding SAS Web Application Server Configuration” topic in the SAS 9.4 Intelligence Platform: Middle-Tier Administration Guide.

2. Append the mapr.login.conf file to the end of the jaas.config file for the SAS Web Application Server. This server is typically identified as SASServer15 in SAS configuration files.

   Here is an example of a command you can run to append the file:

   ```
   cat SAS-configuration-directory/Lev\n   n/Web/Applications/DataLoader/Hadoop/Configs/mapr.login.conf
   >> SAS-configuration-directory/Lev\n   n/Web/WebAppServer/SASServer15_1/conf/jaas.config
   ```

3. Save the jaas.config file.

Perform the following steps to update the setenv file:

1. Display a command prompt on the SAS Web Application Server.

2. Navigate to the folder that contains the setenv file. The typical paths to this file are as follows:

   - UNIX: SAS-configuration-directory/Lev\n   n/Web/WebAppServer/SASServer15_1_n/bin/setenv.sh
     
     *Note:* If you have multiple server instances, make the changes in each setenv.sh file.

   - Windows (two files): SAS-configuration-directory\Lev\n   n/Web\WebAppServer\SASServer15_1_n\conf\wrapper.conf file and the SAS-configuration-directory\Lev\n   n/Web\WebAppServer\SASServer15_1_n\bin\setenv.bat file.
     
     *Note:* If you have multiple instances of the SAS Web Application Server for SAS Data Loader for Hadoop, make the same changes in each of the files.

3. Add the following lines to the setenv file:

   ```
   export MAPR_IMPERSONATION_ENABLED=true
   export MAPR_HOME=/opt/mapr
   ```

4. Save your changes.
5. Restart the web application servers that were updated. For more information about restarting web application servers, see the “Understanding SAS Web Application Server Management” topic in the SAS 9.4 Intelligence Platform: Middle-Tier Administration Guide.

**Enable Spark Jobs to Run on a MapR Cluster**

Perform this update to enable Spark jobs to run on a MapR cluster.

The HCatalog classpath that is collected in the SAS Data Loader inventory.json file must be modified to remove the following entry:

```
/opt/mapr/hadoop/hadoop-2.7.0/share/hadoop/common/lib/*
```

Removing this literal entry (with the asterisk) prevents the collisions between the Jackson JAR file Spark uses and the Jackson JAR that Hadoop requires.

Perform these steps to enable Spark jobs to run on a MapR cluster:

1. Open a command line on the SAS Workspace Server.
2. Access the inventory.json file. A typical installation path might be: SAS-configuration-directory\Levn\HadoopServer\conf\inventory.json.
3. In the Hadoop classpath entry, remove the following entry:

```
/opt/mapr/hadoop/hadoop-2.7.0/share/hadoop/common/lib/*
```
4. Save your changes.
5. Restart the server to make your changes go into effect.

**Return to Your Scenario for the Next Step**

Select one of the following links to return to your scenario.

1. “Scenario 1: First-time Deployment of All SAS Software” on page 6
2. “Scenario 2: Deployment into an Existing SAS Environment That Uses Hadoop” on page 7
3. “Scenario 3: Redeployment After a Hadoop Update” on page 10
4. “Scenario 4: Switch to a Different Hadoop Distribution” on page 11
5. “Scenario 5: Upgrade Software” on page 13

**Backing Up Saved Directives**

Perform this task when you want to back up saved directives. This is a task that you might do after SAS Data Loader for Hadoop has been in operation for a while.

A directive is a set of instructions to create a job in SAS Data Loader for Hadoop. Directives can be saved in SAS folders. The default path for saved directives is SAS Folders > User Folders > userName > My Folder. Use the Deployment Backup and Recovery Tool to back up your SAS folders. For more information about
SAS folders and the Deployment Backup and Recovery Tool, see the *SAS 9.4 Intelligence Platform: System Administration Guide*.

*Note*: In the current release, you cannot use promotion tools, such as the Export SAS Package wizard or the Import SAS Package wizard, to back up saved directives. These promotion wizards are available in the **Folders** tab of SAS Management Console and SAS Data Integration Studio.

---

**Post-Installation Tasks: SAS Data Loader for Hadoop**

**Overview**

After the relevant tasks described in “Post-Installation Tasks: SAS Configuration” on page 30 have been completed, a Data Loader Administrator should perform the tasks in this section. These tasks verify that SAS Data Loader for Hadoop has been deployed correctly.

A Data Loader Administrator is a SAS user who has been assigned to the **Data Loader: Administration** role in SAS Management Console. For more information about Data Loader Administrators, see “Add at Least One SAS User Account to the Data Loader: Administration Role” on page 33.

**Table 3.4  Post-Installation Tasks: SAS Data Loader for Hadoop**

<table>
<thead>
<tr>
<th>Task and Documentation</th>
<th>When Required</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Verify That SAS Data Loader for Hadoop Is Working” on page 40</td>
<td>Perform this task for new deployments or significant updates of SAS Data Loader for Hadoop. Any SAS Data Loader for Hadoop user can perform this task.</td>
</tr>
<tr>
<td>“Synchronize SAS QKB Metadata between Hadoop and the SAS Middle Tier” on page 43</td>
<td>Perform this task each time a QKB is deployed to the Hadoop cluster. A QKB is deployed when SAS Data Loader for Hadoop is first installed and whenever the QKB on the cluster is replaced. Only a Data Loader Administrator can perform this task.</td>
</tr>
</tbody>
</table>

**Verify That SAS Data Loader for Hadoop Is Working**

**Identify Supported Browsers**

The list of web browsers that SAS supports is updated as new products are tested and approved. For the latest information, visit the SAS 9.4 Support for Web Browsers and Plug-ins page at the following URL: [http://support.sas.com/resources/thirdpartysupport/v94/browsers.html](http://support.sas.com/resources/thirdpartysupport/v94/browsers.html)
**Sign in to SAS Data Loader**

To open SAS Data Loader for Hadoop, enter its URL into a supported web browser. If you were not given the URL, and if your site is using the default port number, enter the following URL: http://hostname/SASDataLoader. If your site is not using the default port number, then use this URL: http://hostname:port-number/SASDataLoader, as shown in this example: http://etlhost17.ourco.com:7980/SASDataLoader.

When you open the web application, the Sign In to SAS window is displayed.

In the **User ID** field, enter *domain-name*\user-name. In the **Password** field enter the password for the specified domain. Finally, click **Sign In** to start the web application. The system responds by displaying the SAS Data Loader for Hadoop Directives page.
Note: Your instance of SAS Data Loader for Hadoop displays authorized directives only, as determined by the SAS user account that you used to sign in.

**Try Some Basic Features**

Follow these steps to verify that SAS Data Loader for Hadoop is working.

1. In the SAS Data Loader for Hadoop directives page, click the directive **Browse Tables**.

2. In the **Source Table** task, click a data source icon.

3. In the data source, click a table icon and then click **View Table**.
The application responds by opening the SAS Table Viewer in a new browser tab. You are now viewing a sample of the data in a Hadoop table.

If you like, you can at this point create and run a directive using any table that is available to you. The table data is safe. All directives (except for Delete Rows) access source data on a read-only basis.

4. To learn about user preferences, open the SAS Data Loader for Hadoop browser tab and click and click Configuration.

5. In the Configuration window, the Hadoop Configuration panel displays server connection information. You can change the system defaults for the SQL environment and the preferred run-time target. If your Hadoop cluster uses Spark or Impala, they are probably selected by default. You can click Test Connection at any time.

6. Click General Preferences. The settings in this panel are applied to all of your new directives. In a new directive, you can override these values without changing the default. The field Maximum length of SAS columns has an impact on performance and target table size.

7. Click Profiles to see the performance-related settings for the Profile Data directive. This directive generates reports on Hadoop tables that contain a wealth of information about the structure and quality of data. You can open profile reports in the Saved Profile Reports directive. You can also open profile reports when selecting source and target tables.

Synchronize SAS QKB Metadata between Hadoop and the SAS Middle Tier

Metadata for the QKB on your Hadoop cluster must be synchronized with the SAS middle tier in order to enhance the performance of data quality directives in SAS Data Loader for Hadoop.

Perform the following steps when SAS Data Loader for Hadoop is first installed and whenever the QKB on the cluster is replaced:

1. Sign in to SAS Data Loader for Hadoop as the Data Loader administrator.
2. Click in the top right corner of SAS Data Loader for Hadoop.
3. Select Configuration → QKB Administration.
   When you select the QKB Administration panel, you might receive a notification window. The notification states that QKB locales cannot be found. This message indicates that the QKB metadata on the Hadoop cluster must be published into the SAS middle tier.
4. Click OK to close the notification window, and then click Publish QKB.
   After the QKB publish operation, access to the QKB Administration panel is restored.
Define Data Sources

Overview

SAS Data Loader for Hadoop enables you to move data between Hadoop and other systems. In order to do that, a SAS administrator must register the data sources in these other systems in SAS metadata.

Table 3.5 SAS Data Loader for Hadoop Data Sources

<table>
<thead>
<tr>
<th>Task and Documentation</th>
<th>When Required</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Define JDBC Data Sources” on page 44</td>
<td>Perform this task when you want to support parallel loading of DBMS data in and out of the Hadoop cluster. Using JDBC drivers is optional, but recommended. This task is typically done after the Hadoop administrator installs JDBC drivers on the cluster.</td>
</tr>
<tr>
<td>“Define SAS Libraries” on page 50</td>
<td>Perform this task if you want to copy data between SAS and Hadoop. This task can be done anytime after SAS Data Loader for Hadoop has been deployed.</td>
</tr>
<tr>
<td>“Connect to SAS LASR Analytic Server” on page 50</td>
<td>Perform this task if you want to copy Hadoop columns into memory on SAS LASR Analytic Server for analysis using SAS Visual Analytics or SAS Visual Statistics.</td>
</tr>
<tr>
<td>“Customize an Impala Server JDBC Connection to Support Special Use Cases” on page 52</td>
<td>Perform this task when you want to support one of these use-cases for a JDBC connection to a Cloudera Impala server:  • your Hadoop cluster is secured with LDAP  • your Hadoop cluster is secured with Kerberos, and you want to use the Cloudera Impala driver</td>
</tr>
</tbody>
</table>

Define JDBC Data Sources

Overview

SAS software on the cluster uses the Sqoop and Oozie components installed with the cluster to support parallel loading of DBMS data in and out of the cluster. SAS Data Loader for Hadoop also accesses databases directly, using JDBC for the purpose of selecting either source or target schemas and tables to move.

Using JDBC drivers is optional, but recommended. If you decide to use them, the SAS administrator performs the following tasks to support JDBC data sources:
Asks the Hadoop administrator for a copy of the JDBC drivers that are installed on the Hadoop cluster. Each driver supports a connection to a DBMS, such as Oracle or Teradata.

Installs the JDBC drivers on the SAS Web Application Server on page 45.

Creates a SAS authentication domain for each DBMS to be accessed via JDBC on page 45.

Registers the DBMS servers in SAS Management Console on page 46. The connection metadata for each DBMS includes one of the authentication domains above.

Adds extended attributes to the connection metadata for each DBMS server on page 48. These attributes point to the JDBC driver for that DBMS on the SAS Web Application Server.

Copy JDBC Drivers to the SAS Web Application Server

The Hadoop administrator installs JDBC drivers for one or more DBMSs on the Hadoop cluster. The SAS administrator installs the same drivers on the SAS Web Application Server for SAS Data Loader for Hadoop. The drivers on the Hadoop cluster and the SAS Web Application Server must match exactly.

Perform these steps to obtain JDBC drivers from the Hadoop administrator and copy them to the SAS Web Application Server.

1. Ask your Hadoop administrator for a copy of the JDBC drivers that are installed on your Hadoop cluster. For more information about this task on Hadoop, see “Install and Configure JDBC Drivers on the Hadoop Cluster” on page 74.

2. Open a command prompt on the SAS Web Application Server for SAS Data Loader for Hadoop.

   SAS Data Loader for Hadoop has a dedicated SAS Web Application Server. By default, this server is identified as SASServer15_n in a list of running processes. It is installed in a directory such as SAS-configuration-directory/Lev1/Web/WebAppServer/SASServer15_1. For more information about accessing web application servers, see the “Understanding SAS Web Application Server Configuration” topic in the SAS 9.4 Intelligence Platform: Middle-Tier Administration Guide.

3. Navigate to the JDBC drivers folder. Here is a typical path to this folder:

   \SAS-configuration-directory\Lev1\Web\Applications\DataLoader\Jdbc\Drivers

4. Copy the files for the JDBC drivers into the folder.

5. Restart this server in order for the drivers to be available.

   For more information about restarting web application servers, see the “Understanding SAS Web Application Server Management” topic in the SAS 9.4 Intelligence Platform: Middle-Tier Administration Guide.

The JDBC drivers that were installed on the SAS Web Application Server support one or more DBMSs. The next task is to register the corresponding DBMS servers in SAS Management Console.

Create an Authentication Domain for Each DBMS Accessed via JDBC

JDBC connections do not support single sign-on when the Hadoop cluster is secured with Kerberos. However, there is another way to automate authentication for JDBC
connections. You can create a SAS authentication domain for each DBMS to be accessed via JDBC.

For example, to access an Oracle database with a JDBC connection, you can define a SAS authentication domain for Oracle. You can add logins for SAS Data Loader users or groups to the domain. When you register the Oracle server in SAS Management Console, you can select the Oracle domain. When people use the Oracle server definition to connect to Oracle, if they have a login in the Oracle domain, they can be authenticated automatically.

To create a SAS authentication domain, see “Manage Authentication Domains” in SAS 9.4 Management Console: Guide to Users and Permissions.

To add logins for users or groups to that domain, see “About Logins” in SAS 9.4 Management Console: Guide to Users and Permissions.

**Register DBMS Servers in SAS Management Console**

The JDBC drivers that were installed on the SAS Web Application Server support one or more DBMSs. Use SAS Management Console to register a server for each DBMS. Be sure to select one of the authentication domains that are described in the previous topic. For an example of how to create a DBMS connection, see “Establishing Connectivity to an Oracle Database” in the SAS 9.4 Intelligence Platform: Data Administration Guide.
Figure 3.3 Example SAS Metadata for an Oracle Server

Note the connection metadata in the previous figure (Connection: ORACLE_TARG) . The connection metadata should include one of the authentication domains that are described in the previous topic. An example domain (DMTARG) is shown in the next figure.
The next task is to add extended attributes to the DBMS connection metadata.

**Add Extended Attributes to DBMS Connections**

SAS Data Loader for Hadoop can use the metadata for a DBMS server and the credentials of the current user to generate a default connection string for the server. By specifying the extended attributes below, you can replace the default connection string with one that uses a JDBC driver for that DBMS.

The next table describes some extended attributes that can be used to specify a JDBC connection to a DBMS.

**Table 3.6 Main Extended Attributes for JDBC Connections**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description and Example Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>connection.jdbc.url</td>
<td>URI that specifies a JDBC connection to a DBMS server.</td>
</tr>
<tr>
<td></td>
<td>jdbc:oracle:thin:@ed01-scan.unx.Mycompany.com:1521/exadat</td>
</tr>
<tr>
<td></td>
<td>jdbc:teradata://tera2650.unx.Mycompany.com/Database=dmtarg</td>
</tr>
<tr>
<td>connection.jdbc.driverclassname</td>
<td>Name of the JDBC driver class used to connect to the DBMS server. The same JDBC drivers must be</td>
</tr>
<tr>
<td></td>
<td>on the Hadoop cluster and on the dedicated SAS Web Application Server for SAS Data Loader for</td>
</tr>
<tr>
<td></td>
<td>Hadoop (SASServer15_n).</td>
</tr>
<tr>
<td></td>
<td>oracle.jdbc.OracleDriver</td>
</tr>
<tr>
<td></td>
<td>com.teradata.jdbc.TeraDriver</td>
</tr>
</tbody>
</table>
Perform the following steps to add JDBC attributes to the connection metadata for a DBMS server.

1. Log on to SAS Management Console as an administrator.
2. From the **Plugins** tab, select **Environment Management > Server Manager**.
3. In the left pane, select the metadata object for a DBMS server, such as an Oracle server. If a connection has been defined, the metadata for the connection appears in the right pane.

**Figure 3.5  Connection Metadata for an Oracle Server**

4. Right-click the metadata for the connection in the right pane and select **Properties**.
5. Click the **Extended Attributes** tab. This tab is where you can customize the server connection.

6. To customize the connection, click **New** to add the extended attributes shown in the previous table. The next figure shows an example of the attributes for an Oracle server connection.
Click **OK** to save your changes in SAS Management Console.

Note the `connection.jdbc.property.defaultRowPrefetch` attribute in the previous figure. This is a JDBC property defined by Oracle for Oracle servers. The value for this property is an integer that specifies the default number of rows to prefetch from the server. The default is 10. You might want to review the custom JDBC attributes that are supported by a DBMS.

### Define SAS Libraries

The Copy Data to Hadoop directive can copy data into Hadoop from SAS. SAS data sources include SAS LASR Analytic Servers, SAS data sets, and any data source that can be connected to using a SAS LIBNAME engine. Similarly, the Copy Data from Hadoop directive can copy data from Hadoop to SAS. Defining a SAS library is a standard SAS task. For more information about this task, see the SAS 9.4 Intelligence Platform: Data Administration Guide.

### Connect to SAS LASR Analytic Server

If your site has licensed SAS Visual Analytics or SAS Visual Statistics, SAS Data Loader can copy Hadoop columns into memory on SAS LASR Analytic Server. Once there, the columns can be analyzed. To support this scenario, a SAS administrator specifies a connection to the SAS LASR Analytic Server. For more information about this task, see “Establishing Connectivity to a SAS LASR Analytic Server” in the SAS 9.4 Intelligence Platform: Data Administration Guide.

If you open the **Properties** dialog box for a SAS LASR Analytic Server, and click the **Options** tab, you see a *Single machine server* field. The next figure shows this field.
If the **Single machine server** field is set to **Yes**, SAS Data Loader uses symmetric multiprocessing (SMP) to load data on SAS LASR Analytic Server. Otherwise, it uses massively parallel processing (MPP).

The following usage notes apply to SAS Data Loader for Hadoop:

- In order for the Load to LASR directive to be successful, SAS LASR Analytic Server cannot be configured to truncate incoming character data. To prevent truncation, do not set the DBMAX_TEXT data set option on the source table for the directive.

- To load data from MapR Hadoop to SAS LASR Analytic Server, the **Single machine server** field in the SAS LASR Analytic Server definition must set to **Yes** (symmetric multiprocessing).
Customize an Impala Server JDBC Connection to Support Special Use Cases

When to Customize the Connection for an Impala Server

The SAS Deployment Wizard prompts for the host name and port of a Cloudera Impala server, if one is available on your cluster. The wizard registers the Impala server on the SAS Metadata Server. SAS Data Loader for Hadoop can use the metadata for the Impala server and the credentials of the current user to generate a default connection string for the server. There are at least two use cases where you must update the SAS metadata for an Impala server:

- your Hadoop cluster is secured with LDAP
- your Hadoop cluster is secured with Kerberos, and you want to use the Cloudera Impala driver

Understanding the Metadata for an Impala Server Connection

This topic explains the basics about SAS metadata for an Impala server, as shown in the next figure.

Figure 3.8 SAS Metadata for an Impala Server

Perform the following steps to view the metadata for an Impala server.

1. Log on to SAS Management Console as an administrator.
2. From the Plugins tab, select Environment Management > Server Manager > SAS Data Loader – Impala Server.
3. In the left pane, select the metadata object for **SAS Data Loader – Impala Server**. If a connection has been defined, the metadata for the connection appears in the right pane.

4. Right-click the metadata for the connection in the right pane and select **Properties**.

5. In the properties dialog box, click the **Options** tab. Basic metadata such as the host name and port is available from this tab. Click the **Option** button to see these values.

6. Click the **Extended Attributes** tab. This tab is where you can customize the Impala server connection. To customize the connection, click **New** to add one or more extended attributes. These attributes are described in the following sections.

7. After you add these attributes, click **OK** to save them.

**Customizing an Impala Connection: Custom Connection URL**

SAS Data Loader for Hadoop uses JDBC to connect to a number of servers, including Impala servers. All of these servers can have a custom connection string in their metadata. This string can be specified by adding a custom connection URL to the **Extended Attributes** tab for the server’s connection.

For example, you can add a `connection.jdbc.url` property to the **Extended Attributes** tab for the connection associated with an Impala server, as shown in the next figure.

**Figure 3.9** Connection String for an Unsecured Impala Server on MyImpalaServer.com, Port 21050

When the `connection.jdbc.url` property is specified, SAS Data Loader for Hadoop uses it instead of generating a default connection string.

The connection string can literal values, as shown in the previous figure. It can also contain placeholder variables that resolve at run time. Placeholder variables are defined using the following syntax: `{propertyname}`. That is an open brace followed by the name of a property (variable) followed by the close brace. By convention, the variable names provided by SAS Data Loader for Hadoop are all lowercase characters. The variables in the next table can be resolved for an Impala server connection string.

**Table 3.7** Basic Connection String Variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>host</td>
<td>Host name for the server. This variable resolves only if a host name is specified in the <strong>Options</strong> tab for the connection metadata. If no host name is specified in the metadata, specify a literal host name in the connection string.</td>
</tr>
<tr>
<td>port</td>
<td>text</td>
</tr>
<tr>
<td>Variables</td>
<td>Description</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>user</td>
<td>User name used to log on to the server. The user, password, and passwordenc variables are resolved based on the identity of user who is currently logged on to SAS Data Loader for Hadoop, together with any authentication domain configured for the Impala server.</td>
</tr>
<tr>
<td>password</td>
<td>Password used to log on to the server.</td>
</tr>
<tr>
<td>passwordenc</td>
<td>SAS002 encoded password string. Typically, not used in JDBC connection strings because the JDBC driver does not handle an encoded password.</td>
</tr>
<tr>
<td>authdomain</td>
<td>Authentication domain specified for all servers connected to via JDBC. If the user does not have credentials for the authentication domain specified for the server, the primary identity is used. This is typically the identity the user used to log on to SAS Data Loader for Hadoop.</td>
</tr>
</tbody>
</table>

An example connection string using these parameters could look like this:

```
jdbc:hive2://{host}:{port}/;user={user};password={password}
```

This is the default connection string used by SAS Data Loader for Hadoop when it detects that Impala is protected by LDAP and the Cloudera Impala driver is not present on the classpath. "Not present on the classpath" means the Cloudera JDBC JAR files and any supporting JAR files are not in the JDBC Drivers directory used by SAS Data Loader for Hadoop.

You have at least two ways to address the special use-cases described below:

- add a custom connection URL to the **Extended Attributes** tab for the connection associated with an Impala server, as described in this topic
- add one or more special properties to the **Extended Attributes** tab for the connection associated with an Impala server, as described in the topics below

### Customizing an Impala Connection: When the Hadoop Cluster Is Secured with LDAP

If LDAP is used to secure the Hadoop cluster, you must update the connection associated with an Impala server so that it specifies LDAP. There are two ways to do this. One way is to add a custom connection URL to the **Extended Attributes** tab for the connection associated with an Impala server.

Here is an example custom connection URL when using the Hive driver:

```
jdbc:hive2://{host}:{port}/;user={user};password={password}
```

The fact that this URL is providing a user name and password and does not specify the `auth=noSasl` parameter tells the driver that LDAP is being used.

If the Cloudera Impala driver is used, the connection string for an LDAP-protected cluster must be specified using the `connection.jdbc.url` property. Here is an example custom connection URL when using the Cloudera Impala driver:

```
jdbc:impala://{host}:{port};AuthMech=3;transportMode=binary;UID={user};PWD={password}
```
If LDAP is used to secure the Hadoop cluster, and a custom connection URL is not provided, SAS Data Loader for Hadoop requires a special property be defined in the Extended Attributes tab for an Impala server connection:

```
connection.jdbc.property.ldapauthentication
```

This property must be set to the string `true`. The next figure shows this property in the Extended Attributes tab for an Impala server connection.

**Figure 3.10** connection.jdbc.property.ldapauthentication Property

---

**Customizing Your Impala Connection: When the Cloudera Impala Driver Is Used with Kerberos**

If Kerberos is used to secure the Hadoop cluster, and you want to use the Cloudera Impala driver, you must update the connection associated with an Impala server. There are two ways to do this. One way is to add a custom connection URL to the Extended Attributes tab for the connection associated with an Impala server. Here is an example custom connection URL with literal values for Kerberos attributes:

```
jdbc:impala://{host}:{port};AuthMech=1;KrbRealm=EXAMPLE.COM;KrbHostFQDN=impala.example.com;KrbServiceName=impala
```

If the use case above applies, and a custom connection URL is not provided, SAS Data Loader for Hadoop requires the set of special properties to be defined in the advanced settings for the Impala connection.

**Table 3.8 Kerberos Connection String Variables**

<table>
<thead>
<tr>
<th>Variables</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>connection.jdbc.property.kerberosprincipal</td>
<td>kerberosprincipal is required when using the Hive JDBC driver (the Cloudera Impala driver is not present on the classpath) and kerberosrealm is not defined. If kerberosrealm is defined, but kerberosprincipal is not, the principal string defaults to: <code>impala/_HOST@{VALUE OF KERBEROSREALM}</code> (no brackets).</td>
</tr>
<tr>
<td>connection.jdbc.property.kerberosrealm</td>
<td>kerberosrealm corresponds to the text after the @ in the principal name. If the Cloudera Impala driver is used and a custom connection URL is not provided, kerberosrealm and kerberosdomain must be specified (kerberosservice defaults to impala).</td>
</tr>
<tr>
<td>connection.jdbc.property.kerberosdomain</td>
<td>kerberosdomain corresponds to the fully qualified domain name that appears before the @ in the principal name.</td>
</tr>
</tbody>
</table>
**Variables**

<table>
<thead>
<tr>
<th>Variables</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>connection.jdbc.property.kerberosservice</td>
<td>See note above for kerberosrealm.</td>
</tr>
</tbody>
</table>

The next figure illustrates how values for these variables could be specified in the **Extended Attributes** tab for an Impala server connection.

**Figure 3.11  Values for Kerberos Connection String Variables**

<table>
<thead>
<tr>
<th>SAS Data Loader – Hadoop Server Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>#</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
</tbody>
</table>

The values in the previous figure are for illustration. It is uncommon for kerberosprincipal and the other three properties to be specified together. Typically, you specify one of these values.

As a third alternative, you could create a custom connection URL with variables for Kerberos attributes:

```text
jdbc:impala://{host}:{port};AuthMech=1;KrbRealm={kerberosrealm};KrbHostFQDN={kerberosdomain};KrbServiceName={kerberosservice}
```

Both the variables and the custom connection URL would be specified in the **Extended Attributes** tab for an Impala server connection.

---

**Update SAS Data Loader to Match a New or Updated Hadoop Distribution**

**Overview**

To configure SAS Data Loader for Hadoop to use an updated Hadoop cluster or to use a new Hadoop distribution, you must do the following:

1. Collect the required Hadoop JAR and configuration files and make them available to the SAS client machines. You can do that in one of the following ways.
   - Use the SAS Deployment Manager to collect the files. SAS Deployment Manager, a tool that enables you to perform some administrative and configuration tasks, is included with each SAS software order.
   - Work with the Hadoop administrator to run the Hadoop tracer script to collect the files.

2. Use the SAS Deployment Manager to update SAS configuration files and metadata on the middle tier and server tier to match Hadoop.

3. Use SAS Deployment Manager to deploy the SAS Data Loader for Hadoop web application with the new configuration.
Note: If you have a clustered SAS middle tier, you must manually redeploy the EAR and WAR files on the additional machines. These steps are also described below.

The tasks in this section are part of the following scenarios:
- “Scenario 3: Redeployment After a Hadoop Update” on page 10. For example, the Hadoop cluster is updated with a new parcel, package, service, or component on an existing cluster; or you make any configuration changes to the Hadoop services or components.
- “Scenario 4: Switch to a Different Hadoop Distribution” on page 11. For example, you want to switch from Cloudera to Hortonworks.

Collect the Hadoop JAR and Configuration Files

Using SAS Deployment Manager
Run the SAS Deployment Manager to make the required Hadoop JAR and configuration files available to the SAS client machine.

Note: Alternatively, a Hadoop administrator can run a tracer script at the Hadoop cluster to collect the files. For details, see “Use the Hadoop Tracer Script to Collect Hadoop JAR and Configuration Files” on page 108.

1. Make sure you have the following information or that the following prerequisites are met.
   - For the Hadoop cluster manager:
     - the host name and port number for your Hadoop cluster manager host
     - a UNIX user account with SSH for the cluster manager host
     - credentials for the cluster manager application (Cloudera Manager or Ambari); this is a UNIX account name that has a password and administrator privileges for the cluster manager software
   - Hive service host name and port number
     Note: Ensure that your Hive service is working properly. One way to do that is to issue this command to check if the cluster is responding in a timely manner:
     
     -bash-4.1$ time hive -e 'set -v'
   - Oozie service host name and port number
   - Impala service host name and port number
   - The private key file or the user ID and password for the UNIX user account with SSH access to the machine that is hosting the Hive, Oozie, and Impala services
   - Ensure that Python 2.6 or later and strace are installed. Contact your system administrator if these packages are not installed on the system.
   - For clusters secured with Kerberos, a valid ticket for the user on the client machine and the Hive service
   - For cluster managers that are enabled with Transport Layer Security (TLS), a trusted certificate of authority (CA) must be on the host that you access. You must run the SAS Deployment Manager task **Add Certificate to Trusted CA Bundle** if the trusted CA has not been added to your SAS installation.
Note: All discussion of TLS is also applicable to the predecessor protocol, Secure Sockets Layer (SSL).

• The HDFS user home directory, `/user/user-account`, must exist and have Write permission for the `user-account` or the `mapred` account must have a `drwxrwxrwx` permission for the HDFS/user directory.

Note: This is a critical prerequisite for a MapR cluster.

• Authorization to issue commands for HDFS and Hive. During SAS Deployment Manager processing, a simple validation test is run to see whether HDFS (hadoop) and Hive (hive) commands can be issued. If the validation test fails, the script that pulls the JAR and configuration files is not be executed.

• If you have a Cloudera cluster with Sentry and RecordService services, the Sentry and RecordService services must be configured on the HiveServer2 node. This ensures that the configuration files that are required for Sentry and RecordService services are obtained when SAS Deployment Manager is run.

2. Go to the host running the SAS middle tier.

3. Record the paths where the Hadoop JAR and configuration files are located. For more information about the locations, see “Verify Environment Variables Related to Hadoop” on page 25.

Note: You will navigate to these directories in a later step. Examples of the directories are `.../hadoop/lib` (JAR files) and `.../hadoop/conf` (configuration files).

4. If you are running on a cluster with Kerberos, you must kinit the HDFS user.
   a. Log on to the server using SSH as root with sudo access.
      ```
      ssh username@serverhostname
      sudo su - root
      ```
   b. Enter the following commands to kinit the HDFS user. The default HDFS user is `hdfs`.
      ```
      su - hdfs | hdfs-userid
      kinit -kt location of keytab file
      user for which you are requesting a ticket
      ```

Note: For all Hadoop distributions except MapR, the default HDFS user is `hdfs`. For MapR distributions, the default HDFS user is `mapr`.

Note: If you are running on a cluster with Kerberos, a valid keytab is required for the HDFS user who configures the Hadoop JAR and configuration files.

To check the status of your Kerberos ticket on the server, run `klist` while you are running as the -hdfsuser user. Here is an example:

```
klist
Ticket cache: FILE/tmp/krb5cc_493
Default principal: hdfs@HOST.COMPANY.COM

Valid starting   Expires       Service principal
06/20/16 09:51:26 06/27/16 09:51:26 krbtgt/HOST.COMPANY.COM@HOST.COMPANY.COM
renew until 06/27/16 09:51:26
```

5. Start SAS Deployment Manager by running `sasdm.exe` for Windows or `sasdm.sh` for UNIX. The SAS Deployment Manager script is located in the `/SASHome/SASDeploymentManager/9.4` directory.

Perform one of the following steps to start the SAS Deployment Manager:
Update SAS Data Loader to Match a New or Updated Hadoop Distribution

- UNIX: Navigate to $SAS$-installation-directory/SASDeploymentManager/9.4, and then run ./sasdm.sh.
- Windows: Start Windows Explorer, and then navigate to $SAS$-installation-directory\SASDeploymentManager\9.4. Next, right-click sasdm.exe, and then select Run as administrator.

Note: Run the SAS Deployment Manager as the user who performed the original installation of SAS Data Loader for Hadoop.

The Choose Language page opens.

6. Select the language that you want to use to perform the configuration of your software.

   Click OK. The Select SAS Deployment Manager Task page opens. The items listed under Hadoop Configuration depend on the SAS software that you have licensed.

7. Under Hadoop Configuration, select Configure Hadoop Client Files.

   Click Next. The Select Hadoop Distribution page opens.

8. From the drop-down menu, select the distribution of Hadoop that you are using.

   Note: If you are switching to a different Hadoop distribution, select the target Hadoop distribution.

   Note: If your MapR client is on Windows, the MAPR_HOME and JAVA_HOME environment variables must be set. For more information, see MapR: Setting Up the Client.

   Click Next.

   If your Hadoop distribution does not have an administrative client, the Hadoop Cluster Service Information page opens. Skip to Step 12 on page 60.

   If your Hadoop distribution has an administrative client such as Cloudera Manager or Ambari, the Use Cluster Manager page opens.

9. Select the cluster manager administrative tool from the list.

   The Hive and Oozie services information that SAS Deployment Manager needs to configure the Hadoop client files can be retrieved from the cluster manager. Select the cluster manager that you want to use to retrieve the information, or select None if you want to specify the information yourself.

   If you select None and click Next, the Hadoop Cluster Service Information page opens. Skip to Step 12 on page 60.

   For MapR, browse to enter the MapR client directory, and click Next. The Hadoop Cluster Service Information page opens. Skip to Step 12 on page 60.

   If you select a cluster manager and click Next, the Hadoop Cluster Manager Information page opens.

10. Enter the following information:

   - Enter the host name and port number for your Hadoop cluster manager.

     For Cloudera, enter the location where Cloudera Manager is running. For Hortonworks, IBM BigInsights, or Pivotal, enter the location where the Ambari server is running.

     The port number is set to the appropriate default after Cloudera, Hortonworks, IBM BigInsights, or Pivotal is selected above.
Note: The host name must be a fully qualified domain name. The port number must be valid, and the cluster manager must be listening. The SAS Deployment Manager validates a constructed URL based on the host name and port number and issues an error message if the URL is not accessible.

- Indicate whether the cluster manager is enabled with Transport Layer Security (TLS).

Note: All discussion of TLS is also applicable to the predecessor protocol, Secure Sockets Layer (SSL).

Note: If you choose Yes and a trusted certificate authority (CA) is not found for the host that you are trying to access, an error occurs. A dialog box that prompts you to run the SAS Deployment Manager Add Certificate to Trusted CA Bundle task is displayed. For more information about how to run this task, see “Add Certificate to Trusted CA Bundle” in SAS Deployment Wizard and SAS Deployment Manager: User’s Guide. After you run this task to add the certificate, you must restart the Configure Hadoop Client Files task.

Click Next. The Hadoop Cluster Manager Credentials page opens.

11. Enter the Cloudera Manager or Ambari administrator account name and password.

Click Next. The Hadoop Cluster Service Information page opens.

12. Enter the following information:

- The host names of the Hive, Impala, and Oozie services for the Hadoop cluster. If you use the cluster manager, this field is populated for you.

Note: If your SAS software uses Impala or Oozie, enter the Impala or Oozie service host name so that the correct JAR and configuration files are collected. In addition, the host names for Impala and Impala are added to the inventory.json file that SAS Deployment Manager creates in this step.

- The method for collecting Hadoop JAR and configuration files. Choose one of these three methods:

  - To use a password of the UNIX user account with SSH for the Hive, Impala, and Oozie hosts, select SAS Deployment Manager with the UNIX Account with SSH - Password.

Click Next. The Hadoop Cluster SSH Credentials page opens and asks for a password. Continue with Step 13 on page 61.

  - To use a private key file (id_rsa) of the UNIX user account with SSH for the Hive, Impala, and Oozie hosts, select SAS Deployment Manager with the UNIX Account with SSH - Private Key Files.

Click Next. The Hadoop Cluster SSH Credentials page opens and asks for a private key file. Skip to Step 14 on page 61.

  - If you do not have SSH credentials or want to manually collect the Hadoop JAR and configuration files, select Manual Collection, and follow the instructions in the dialog box that displays.

Note: The Hadoop JAR and configuration files must be collected before proceeding with the SAS Deployment Manager task. The configuration will fail if the Hadoop JAR and configuration files are not available during the remaining configuration tasks that are executed by SAS Deployment Manager. For information about how to manually collect the Hadoop JAR and configuration files, see “Use the Hadoop Tracer Script to Collect Hadoop JAR and Configuration Files” on page 108.
After you collect the files using the Hadoop tracer script, click Next on the Hadoop Cluster Service Information page. The Hadoop Cluster Service Port Information page opens, and skip to Step 15 on page 61 to continue.

13. Enter the account name and password of the UNIX user who has SSH access to the machine that is hosting the Hive, Impala, and Oozie services. This information is required to move and copy files to and from hosts.

   **Note:** If Kerberos is installed on your Hadoop cluster, then the user should have a Kerberos principal configured.

   Click Next. The Hadoop Cluster Service Port Information page opens. Skip to Step 15 on page 61.

14. Enter the following information. This information is required to move and copy files to and from hosts.

   • The account name of the UNIX user who has SSH access to the machine that is hosting HiveServer2.

   • The path to the location of the private key file.

   **Note:** If Kerberos is installed on your Hadoop cluster, then the user should have a Kerberos principal configured.

   Click Next. The Hadoop Cluster Service Port Information page opens.

15. Enter the port numbers of the Hive, Impala, and Oozie services of your Hadoop cluster.

   Click Next. The Specify SAS Hadoop Client Directories page opens.

16. Navigate to the enter the location of the Hadoop configuration files and JAR files that you recorded in Step 3 on page 58.

   **TIP** It is recommended that you select Most recent version available to replace duplicate JAR files by the latest version.

   **Note:** For MapR distributions that need to use Pig, set Filter JAR Files By to No filter. This setting enables the correct JAR files to be pulled.

   **CAUTION:**
   The configuration files and JAR files for the Hadoop client must reside in the /conf and /lib directories, respectively. You can specify a non-default path to the /conf and /lib directories. If you do not have the /conf and /lib directories, SAS software cannot find the required files to run successfully.

   Click Next. The Update SAS Configuration File sasv9.cfg page opens.

17. DO NOT SELECT the add Add environment variables check box. If the check box is selected, be sure to deselect it.

   Click Next.

18. (Optional) Validate the configuration of SAS/ACCESS Interface to Hadoop.

   If you want to collect the JAR and configuration files without validation, deselect this option.

   If there are problems with the validation, an error message appears. You can check the log file checkaccesshdp_timestamp.log, for the cause of the error. By default, the validation log file can be found in your account home directory:

   • UNIX: /your-home/.SASAppData/SASDeploymentWizard

   • Windows: C:\users\your-account\AppData\Local\SASDeploymentWizard
Click Next. The **Hadoop Cluster Hive Service Information** page opens.

19. Enter the schema name for the cluster’s Hive service and select whether Kerberos is enabled on the cluster.

A valid Kerberos ticket must be available on the client machine and Hive service. If a ticket is not available, you must go out to the client machine, cluster, or both and obtain the Kerberos ticket. When the ticket is obtained, you can resume the deployment using SAS Deployment Manager.

**Note:** If you are using Advanced Encryption Standard (AES) encryption with Kerberos, manually add the Java Cryptography Extension `local_policy.jar` file in every location where JAVA Home resides on the cluster. If you are located outside the United States, you must also manually add the `US_export_policy.jar` file. The addition of these files is governed by the United States import control restrictions. These two JAR files also need to replace the existing `local_policy.jar` and `US_export_policy.jar` files in the SAS JRE location, which is the `SASHome/SASPrivateJavaRuntimeEnvironment/9.4/jre/lib/security/` directory. As a best practice, back up the existing `local_policy.jar` and `US_export_policy.jar` files first in case they need to be restored. These files can be obtained from the IBM or Oracle websites.

Click Next. The **Hive User Name Credentials** page opens.

20. Enter a Hive user name that has access to the Hive service and the password for that Hive user name.

**Note:** Check with your Hadoop administrator for a valid user name and password.

Click Next. SAS Deployment Manager verifies the prerequisites for the validation and checks for locked files and Write permissions. Checking the system might take several seconds. The **Checking System** page opens.

21. If any files are shown in the text box after the system check, follow the instructions on the **Checking System** page to fix any problems.

Click Next. The **Summary** page opens.

22. Click **Start** to begin the configuration.

**Note:** It takes several minutes to complete the configuration. If Kerberos is installed on your Hadoop cluster, the configuration could take longer.

If the configuration is successful, the name of the page changes to **Deployment Complete** and a green check mark is displayed beside SAS Hadoop Configuration Libraries.

**Note:** Part of the configuration process runs SAS code to validate the environment. A green check mark indicates that SAS Deployment Manager could connect to Hadoop, run a tracer script, pull back files, and run SAS code to validate the setup.

If warnings or errors occur, fix the issues and restart the configuration.

23. Click **Next** to close SAS Deployment Manager.

24. Proceed to “Update SAS Metadata for Hadoop Client Files” on page 63.

**Using the Hadoop Tracer Script**

If you do not have access to the SAS Deployment Manager, perform the following steps to collect the Hadoop JAR and configuration files.
1. Work with the Hadoop administrator to run the Hadoop tracer script to collect the files. For details, see “Use the Hadoop Tracer Script to Collect Hadoop JAR and Configuration Files” on page 108.

2. Back up the original Hadoop JAR and configuration (XML) files on the middle tier and compute tier. Here are the typical paths to the files.

   On the middle tier:
   - UNIX:
     - `SAS-configuration-directory/Lev'n/Web/Applications/DataLoader/Hadoop/Configs/`
     - `SAS-configuration-directory/Lev'n/Web/Applications/DataLoader/Hadoop/Jars/`
   - Windows:
     - `SAS-configuration-directory\Lev'n\Web\Applications\DataLoader\Hadoop\Configs\`
     - `SAS-configuration-directory\Lev'n\Web\Applications\DataLoader\Hadoop\Jars\`

   On the compute tier:
   - UNIX:
     - `SAS-configuration-directory/Lev'n/HadoopServer/conf/`
     - `SAS-configuration-directory/Lev'n/HadoopServer/lib/`
   - Windows:
     - `SAS-configuration-directory\Lev'n\HadoopServer\conf\`
     - `SAS-configuration-directory\Lev'n\HadoopServer\lib\`

3. Copy the files collected by the Hadoop tracer script to the respective paths shown in Step 2 on page 63.

   **TIP** On the compute tier, the JAR files are placed in the `lib` directory.

4. Proceed to “Update SAS Metadata for Hadoop Client Files” on page 63.

---

**Update SAS Metadata for Hadoop Client Files**

The values in the SAS configuration files for Hadoop are registered in SAS metadata. If you update the configuration files, you must update the metadata to match.

1. Go to the SAS server on the SAS middle tier.

2. Perform one of the following steps to start the SAS Deployment Manager:
   - UNIX: Navigate to `SAS-installation-directory/\SASDeploymentManager/9.4`, and then run `./sasdm.sh`.
   - Windows: Start Windows Explorer, and then navigate to `SAS-installation-directory\SASDeploymentManager\9.4`. Next, right-click `sasdm.exe`, and then select **Run as administrator**.

3. Scroll to the list of tasks under Hadoop Configuration.

4. Select **Update Metadata Configuration for Hadoop Client Files**, and click **Next**.

5. Select the configuration directory and level, and click **Next**.
6. Enter the password for the SAS Metadata Server, and click Next.

7. Perform the following steps:
   a. Locate the `sas_hadoop_config.properties` file. Typically, this is at the same level as the `sashome` configuration directory, in a subdirectory named `hadoop`.
   b. Type in the complete path and filename of the `sas_hadoop_config.properties` file, and click Next.

8. Click Start to begin the configuration.

9. Click Next when the configuration is completed, and then click Finish to exit the SAS Deployment Manager.

10. Proceed to “Deploy the SAS Data Loader for Hadoop Web Application” on page 64.

**Deploy the SAS Data Loader for Hadoop Web Application**

After you update the SAS metadata to match the updated SAS configuration files for Hadoop, you must deploy the SAS Data Loader for Hadoop web application. This manually deploys the updated WAR and EAR files for the application.

Perform these steps:

1. Verify that you have the password for the SAS Metadata Server.

2. Go to a SAS server on the middle tier.

3. Perform one of the following steps to start the SAS Deployment Manager:
   - UNIX: Navigate to `SAS-installation-directory/SASDeploymentManager/9.4`, and then run `./sasdm.sh`.
   - Windows: Start Windows Explorer, and then navigate to `SAS-installation-directory\SASDeploymentManager\9.4`. Next, right-click `sasdm.exe`, and then select Run as administrator.

4. Scroll to the list of tasks under Administrative Tasks.

5. Select Deploy Web Applications, and click Next.

6. Select the configuration directory and level, and click Next.

7. Enter the password for the SAS Metadata Server, and click Next.

8. Select the Allow the application server to stop check box, and click Next.

9. Select the Data Loader Mid-Tier check box, and click Next.

10. Click Start to begin the configuration.

11. Click Next when the configuration is completed, and then click Finish to exit the SAS Deployment Manager.

**Additional Steps for Horizontal Clusters**

If you have a clustered SAS middle tier, you must manually redeploy the web applications on the additional machines.
• Generally, perform the steps described in the Additional Steps for Horizontal Clusters topic. This topic is part of the instructions for redeploying SAS web applications in the SAS 9.4 Intelligence Platform: Middle-Tier Administration Guide.

• When you copy the updated EAR and WAR files from the primary machine to the staging directory on the additional machines, copy the files located in SAS-configuration-directory/Levn/Web/Applications/DataLoader directory.

Return to Your Scenario for the Next Step

Select one of the following links to return to your scenario.

- “Scenario 1: First-time Deployment of All SAS Software” on page 6
- “Scenario 2: Deployment into an Existing SAS Environment That Uses Hadoop” on page 7
- “Scenario 3: Redeployment After a Hadoop Update” on page 10
- “Scenario 4: Switch to a Different Hadoop Distribution” on page 11
- “Scenario 5: Upgrade Software” on page 13

Update SAS Data Loader Dependencies

Before You Begin

A SAS administrator performs this task after installing SAS Data Loader for Hadoop into an existing environment where SAS software is using Hadoop. For example, a site might have SAS Data Integration Studio and supporting software installed, and you want to add SAS Data Loader for Hadoop to the installation. For details about the installation steps, see “Scenario 2: Deployment into an Existing SAS Environment That Uses Hadoop” on page 7.

After you install SAS Data Loader for Hadoop into an existing environment where SAS software is using Hadoop, you have two sets of SAS metadata for Hadoop servers: the original set from the existing installation and a new set from SAS Data Loader for Hadoop. For example, the existing metadata for the Hadoop server might be called SAS - Hadoop Server. The new metadata for the Hadoop server might be called SAS Data Loader - Hadoop Server.

The metadata for SAS Data Loader Dependencies will point to the new metadata for the Hadoop servers (such as SAS Data Loader - Hadoop Server). In order to preserve lineage and other important relationships across SAS applications, you must update the metadata for SAS Data Loader Dependencies. Update this metadata so that it points to the original Hadoop server metadata from the existing installation (such as SAS - Hadoop Server).
Identify the Two Sets of Metadata for Hadoop Servers

Your first task is to identify the two sets of SAS metadata for Hadoop servers. Perform the following steps to review the names of the metadata objects for the Hadoop server, the Oozie server, and the Impala server:

1. Log on to SAS Management Console as an administrator.
2. From the Plugins tab, select: Environment Management > Server Manager.
3. Review the names of the metadata objects for the Hadoop server, the Oozie server, and the Impala server.
4. Identify the two sets of SAS metadata for Hadoop servers: the original set from the existing installation and a new set from SAS Data Loader for Hadoop.

Update the Metadata for SAS Data Loader Dependencies

Update the metadata for SAS Data Loader Dependencies so that it points to the original Hadoop server metadata from the existing installation. Perform the following steps:

1. Log on to SAS Management Console as an administrator.
2. From the Plugins tab, select: Application Management > Configuration Manager > Data Loader Midtier Shared Services.
3. Right-click Data Loader Midtier Shared Services and select Properties.
4. Click the Settings tab.
5. Select SAS Data Loader Dependencies in the left pane.
6. Note the values from SAS Data Loader for Hadoop in the Hive server field, the Oozie server field, and the Impala server field.
7. Change the name in the Hive server field to the name of the original Hadoop server metadata from the existing installation. For example, you might change the Hive server name from SAS Data Loader - Hadoop Server to SAS - Hadoop Server.
8. Change the name in the Oozie server field to the name of the original Oozie server metadata from the existing installation.
9. Change the name in the Impala server field to the name of the original Impala server metadata from the existing installation.
10. Click OK to save your changes in SAS Management Console.

SAS Data Loader for Hadoop has a dedicated SAS Web Application Server. By default, this server is identified as SASServer15_n in a list of running processes. It is installed in a directory such as SAS-configuration-directory\Lev1\Web\WebAppServer\SASServer15_1. For more information about restarting these servers, see the “Understanding SAS Web Application Server Management” topic in the SAS 9.4 Intelligence Platform: Middle-Tier Administration Guide.
Optional: Delete the Unused Metadata for Hadoop Servers

After you make the changes above, you no longer need the SAS Data Loader for Hadoop metadata objects for Hadoop servers. Perform the following steps to delete these objects:

1. From the Plugins tab, select: **Environment Management > Server Manager**.
2. Identify the names of the metadata objects for the Hadoop server, the Oozie server, and the Impala server that were installed with SAS Data Loader for Hadoop.
3. Delete these objects.

Configure Impala

Before You Begin

A SAS administrator performs this task when Impala is installed on a Hadoop cluster, but it is not registered in SAS metadata and associated with SAS Data Loader for Hadoop.

In most cases, if Impala is installed on a Hadoop cluster, it is registered in SAS metadata when SAS Data Loader for Hadoop is installed. However, there are scenarios where Impala is not automatically registered. For example, suppose that your site switched from one Hadoop distribution to another. If Impala was not installed with the old Hadoop distribution, but it is installed on the new distribution, Impala is not automatically registered. In that case, perform the tasks described in this topic:

- “Register the Impala Server in SAS Management Console” on page 67
- Specify the Impala server in SAS Data Loader Dependencies on page 68

Register the Impala Server in SAS Management Console

For an example of how to register an Impala server in SAS Management Console, see “Establishing Connectivity to an Impala Server” in the SAS 9.4 Intelligence Platform: Data Administration Guide.

*Note:* Write down the exact name of the metadata object for the Impala server. This name will be used to associate this server with SAS Data Loader for Hadoop.
For example, the next figure shows a metadata object called **Impala Server**.

*Figure 3.12  Metadata Object for an Impala Server*

---

**Specify the Impala Server in SAS Data Loader Dependencies**

Update the metadata for **SAS Data Loader Dependencies** so that it points to the Impala server that you registered in the previous topic. Perform the following steps:

1. Log on to SAS Management Console as an administrator.
2. From the **Plugins** tab, select: **Application Management > Configuration Manager > Data Loader Midtier Shared Services**.
3. Right-click **Data Loader Midtier Shared Services** and select **Properties**.
4. Click the **Settings** tab.
5. Select **SAS Data Loader Dependencies** in the left pane.
6. In the **Impala server** field, enter the exact name of the metadata object for the Impala server. For example, if the name of the metadata object is **Impala Server**, you would specify that name.
7. Click **OK** to save your changes in SAS Management Console.

SAS Data Loader for Hadoop has a dedicated SAS Web Application Server. By default, this server is identified as `SASServer15_n` in a list of running processes. It is installed in a directory such as `SAS-configuration-directory\Level\Web\WebAppServer\SASServer15_1`. For more information about restarting these servers, see the “Understanding SAS Web Application Server Management” topic in the SAS 9.4 Intelligence Platform: Middle-Tier Administration Guide.
Chapter 4
Tasks for the Hadoop Administrator

Hadoop Administrator Checklist ............................................................... 72
Get Files and Information from the SAS Administrator ......................... 72
Verify Supported Hadoop Distributions ............................................... 72
Verify General Knowledge of Hadoop ............................................... 73
Install and Configure JDBC Drivers on the Hadoop Cluster .................... 74
Set Up UNIX User Accounts and User Home Directories ...................... 75
About Active Directory (LDAP) Authentication ................................ 76
Tasks for MapR .......................................................... 76
Return to Your Scenario for the Next Step ....................................... 77

Kerberos Administrator Checklist ......................................................... 77
Verify Supported Hadoop Distributions ............................................... 77
Verify Kerberos Tickets and Settings ............................................... 77
Verify Configuration for MIT Kerberos ........................................... 78
Additional Resources and Information ............................................. 78
Return to Your Scenario for the Next Step ....................................... 78

Deploy the SAS Embedded Process .................................................... 79
Before You Begin .......................................................... 79
Using Cloudera Manager ......................................................... 79
Using Ambari .......................................................... 81
Using the SASEP-ADMIN.SH Script ........................................ 82
Return to Your Scenario for the Next Step ....................................... 92

Post-Installation Tasks for the SAS Embedded Process .............................. 93
Before You Begin .......................................................... 93
Additional Configuration Needed to Use HCatalog File Formats .......... 93
MapR: Adding the YARN Application CLASSPATH to the Configuration File 94
(Optional) Changing the Trace Level ........................................ 95
(Optional) Adjusting the SAS Embedded Process Performance .............. 95
Return to Your Scenario for the Next Step ....................................... 98

Deploy the SAS Data Loader for Hadoop Spark Engine ............................. 98
Before You Begin .......................................................... 98
Using Cloudera Manager ......................................................... 98
Using Ambari .......................................................... 100
Using the SASDMP_ADMIN.SH Script .................................... 101
Return to Your Scenario for the Next Step ....................................... 107

Use the Hadoop Tracer Script to Collect Hadoop JAR and Configuration Files . 108
Overview .......................................................... 108
Prerequisites .......................................................... 108
Run the Hadoop Tracer Script ..................................................... 109
Hadoop Administrator Checklist

Refer to the following checklist to ensure that the Hadoop cluster that SAS Data Loader for Hadoop will connect to is ready.

Get Files and Information from the SAS Administrator

The following software must be deployed to the Hadoop cluster before SAS Data Loader for Hadoop is installed.

SAS Embedded Process

The required file and documentation to deploy the SAS Embedded Process are different by release.

| Table 4.1  SAS Embedded Process: File and Documentation by Release |
|---|---|
| SAS Release | File and Documentation |
| SAS 9.4M6 release (November 2018) | • File: SAS_Viya_deployment_data.zip  
• Documentation: To deploy, configure, upgrade, and uninstall the SAS Embedded Process, see the SAS Embedded Process: Deployment Guide.  
This release of the SAS Embedded Process is delivered from a SAS Viya repository, so the ZIP file name references SAS Viya. |
| SAS 9.4M5 release (September 2017) | • File: SAS-Software-Depot\standalone_installs\SAS_Core_Embedded_Process_Package_for_Hadoop\13_0\Hadoop_on_Linux_x64\en_sasexe.zip  
• Documentation: Follow the instructions in this guide. |
**Verify Supported Hadoop Distributions**

Review the [SAS 9.4 Supported Hadoop Distributions page](#) to make sure that SAS Data Loader for Hadoop supports your version.

**Verify General Knowledge of Hadoop**

- You need a working knowledge of the Hadoop Distributed File System (HDFS), MapReduce 1, MapReduce 2, YARN, Hive, and HiveServer2 services. For more information, see the Apache website or the vendor’s website.

- Ensure that the HCatalog, HDFS, Hive, MapReduce2, Oozie, and YARN services are running on the Hadoop cluster.

  **Note:** Other services might be required for other SAS products.

- Ensure that Oozie 4.0 or later is installed.

- Your Hadoop cluster must be configured to use Oozie scripts.

- Know the location of the MapReduce2 home.

- Know the host name of the Hive server and the name of the NameNode.

- Determine where the HDFS and Hive servers are running. If the Hive server is not running on the same machine as the NameNode, record the server and port number of the Hive server for future configuration.

- Request permission to restart the YARN service.

- Verify that you can run a MapReduce job successfully.

- Understand and verify your Hadoop user authentication.

- Verify that you can connect to your Hadoop cluster (HDFS and Hive) from your client machine outside of the SAS environment with your defined security protocol.

- Understand how the Hadoop cluster is secured, such as by Kerberos and Sentry.

- To support Sqoop, verify that the Oozie and Sqoop groups are added to the Hive `proxyuser` list.
Install and Configure JDBC Drivers on the Hadoop Cluster

Why Use JDBC Drivers?
SAS Data Loader for Hadoop uses Sqoop and Oozie to move data to and from a DBMS. It can also use JDBC connections to access a DBMS directly, and to select source tables, target tables, and target schemas.

Note: Using JDBC connections is optional for SAS Data Loader for Hadoop, but it is recommended. These connections can improve performance when moving data to and from a DBMS.

Steps for Installing and Configuring the Drivers
To install and configure JDBC drivers on the Hadoop cluster, perform the following steps:

1. Work with the SAS administrator to identify the DBMSs that SAS Data Loader for Hadoop will access. This can help you decide which JDBC drivers should be installed.
2. Follow the JDBC vendor instructions to download and install JDBC drivers on the Hadoop cluster.
   For Teradata and Oracle, SAS recommends that you download the following JDBC files from the vendor site:
   - Oracle: ojdbc6.jar
   - Teradata: tdgssconfig.jar and terajdbc4.jar. Also, download the Teradata connector JAR file that is matched to your cluster distribution, if it is available.

   Note: Other DBMSs are supported as well.
3. Make sure the JDBC drivers and connector JAR files are located in the Oozie shared libs directory in HDFS, not in /var/lib/sqoop.
   
   Tip: The commands below must be run as the Oozie user or another user with Oozie administrative privilege.
   
   Tip: If Kerberos is used, you must have a valid Kerberos ticket for the Oozie administrative principal before running the commands below.
   
   a. Run the following command to find the location of the JDBC drivers:
      
      oozie admin -oozie http://oozie-server-host:oozie-server-port/oozie -shareliblist sqoop
   
   b. If necessary, copy the JDBC drivers to the Oozie shared libs directory in HDFS.

      Note: You must have, at a minimum, -rw-r--r-- permissions on the JDBC drivers.

4. After JDBC drivers have been copied to HDFS, run the following command to refresh the Oozie Sqoop share lib so that Oozie can recognize the newly added JAR files:

      oozie admin -oozie http://oozie-server-host:oozie-server-port/oozie -sharelibupdate

5. Verify that the JDBC drivers are in the share lib by running the following command:

      oozie admin -oozie http://oozie-server-host:oozie-server-port/oozie -shareliblist sqoop

   Note: After SAS Data Loader for Hadoop has been deployed, the SAS administrator will ask for a copy of the JDBC drivers that are installed on the Hadoop cluster. The
SAS administrator will install the same JDBC drivers on the SAS Web Application Server on the SAS middle tier.

**Set Up UNIX User Accounts and User Home Directories**

Perform the following steps:

1. Create individual UNIX user IDs for each user on all nodes of the Hadoop cluster, as needed, or use a directory service.

   * Note:
     - SAS recommends that all SAS Data Loader for Hadoop users be able to log on to the operating system of the Hadoop cluster.
     - For MapR:
       - All SAS Data Loader for Hadoop users are required to log on to the operating system of the Hadoop cluster. Also, users defined on the cluster must have the same UID and GID as users on the SAS server tier and SAS middle tier.
       - A MapR super user, typically `mapr`, is required on the SAS middle tier with the same GID and UID as defined on the cluster. Share this information with the SAS administrator as this user must start the SAS middle tier.

2. Create a user home directory and Hadoop staging directory in HDFS.

   - The user home directory is `/user/my-user`.
   - The Hadoop staging directory is controlled by the setting `yarn.app.mapreduce.am.staging-dir` in `mapred-site.xml` and defaults to `/user/my-user`.

3. Change the permissions and owner of `/user/my-user` to match the UNIX user.

   - The user must have Read, Write, and Delete permission for the root directory of the Hive warehouse. Examples of this directory are: `/usr/hive/warehouse` and `/apps/hive/warehouse`.
   - The user must have Read, Write, and Delete permission to the HDFS temp directory. Typically, the default is `/tmp`. If the default `/tmp` directory is used, no additional action is required by the SAS administrator.

   * Note: If you do not use the default directory, identify a location that all users need access to, and provide that information to the SAS administrator. The SAS administrator must enter that information into the Oozie server definition in SAS metadata. All users must have Read, Write, and Delete permission to this directory.

   - The Hive service user must have access to the `/user/my-user` directory. Here is an example of the HDFS command to enable this:

     ```
     hdfs dfs -setfacl -m -R default:user:hive:rwx /user
     ```

   - If you use Sentry for authorization, ensure that the following is true:
     - The user has been granted the appropriate Sentry access to the user home directory, the HDFS temp directory, and the default Hive database.
     - The user belongs to a group that is in the `hadoop.proxyuser.hive.groups` list.
About Active Directory (LDAP) Authentication

Oozie does not support LDAP authentication. SAS Data Loader for Hadoop can be enabled to use JDBC connectors and Sqoop and Oozie to move data to and from a DBMS. Accordingly, when LDAP authentication is used with your Hadoop cluster, some SAS Data Loader for Hadoop directives might not work, such as Copy Data To Hadoop using a JDBC access. However, the operation of these directives is otherwise unaffected.

Tasks for MapR

Install the MapR Client

Generally, see the MapR documentation for your release about how to set up the client. Example documentation about this topic can be found here: MapR: Setting Up the Client.

The following items are specific considerations for the SAS environment.

• The version of the MapR client must match the version of the MapR cluster that SAS connects to.

• Consult with the SAS administrator about installing the MapR client on the SAS tiers. The MapR client must be installed on both the SAS server tier where the SAS Workspace Server is located, and the dedicated SAS Web Application Server for SAS Data Loader for Hadoop. By default, the dedicated SAS Web Application Server is identified as SASServer15_n in a list of running processes.

Add SAS Data Loader Users to MapR

Add all SAS Data Loader for Hadoop users to your MapR cluster. See the MapR documentation for your release. Here is some example documentation about this topic: Managing Users and Groups.

Note: Users defined on the cluster must have the same UID and GID as users on the SAS server tier and SAS middle tier.

Enable Impersonation for the MapR Superuser

Enable impersonation for the MapR superuser for your cluster. See the MapR documentation for your release.

Note: SAS Data Loader for Hadoop supports impersonation for the MapR superuser only.

Inform the SAS administrator that the SAS Web Application Server must be configured to support MapR impersonation. This task is described in “Configure the SAS Web Application Server for MapR” on page 38.

Ensure That Oozie Can Execute Spark on a MapR Cluster

SAS Data Loader for Hadoop launches Spark with an Oozie shell command. This command requires the Spark-submit executable to be on the PATH of Oozie’s MapReduce launcher job. For MapR, you must manually add the Spark bin directory to the PATH for Oozie. This can be done by updating the yarn-env.sh script on all nodes in the cluster with a path to the Spark-submit executable.
Perform the following steps to update the `yarn-env.sh` script.

1. Log on to a cluster node.
2. Navigate to the directory that contains the `yarn-env.sh` script. A typical path is: 
   ```bash
   etc/hadoop/yarn-env.sh
   ```
3. Add a line such as the following to the script:
   ```bash
   export PATH=$PATH:/opt/mapr/spark/spark-1.6.1/bin
   ```
4. Apply this change to the `yarn-env.sh` script on all nodes in the cluster.
5. Restart the node manager on all nodes in the cluster. Consult your MapR documentation for details.

---

**Return to Your Scenario for the Next Step**

Select one of the following links to return to your scenario.

- “Scenario 1: First-time Deployment of All SAS Software” on page 6
- “Scenario 2: Deployment into an Existing SAS Environment That Uses Hadoop” on page 7
- “Scenario 3: Redeployment After a Hadoop Update” on page 10
- “Scenario 4: Switch to a Different Hadoop Distribution” on page 11
- “Scenario 5: Upgrade Software” on page 13

---

**Kerberos Administrator Checklist**

Refer to the following checklist to ensure that SAS Data Loader for Hadoop is configured correctly to use Kerberos security.

**Verify Supported Hadoop Distributions**

Review the [SAS 9.4 Supported Hadoop Distributions page](#) to make sure that SAS Data Loader for Hadoop supports Kerberos with your version of Hadoop.

**Verify Kerberos Tickets and Settings**

- To use Kerberos, you must have all valid tickets in place on the cluster. When SAS software on the cluster is deployed, the HDFS user, typically `hdfs`, must have a valid ticket.
- Map each SAS Data Loader for Hadoop to a user principal.
- If you set a maximum lifetime for Kerberos tickets, ensure that the person deploying SAS software on the cluster is aware of the expiration date of the ticket.
- Ensure that the following setting is correct on your cluster:
* hive.server2.enable.doAs = true

**Verify Configuration for MIT Kerberos**

If your cluster is on a different Kerberos realm or Active Directory domain than your SAS Data Loader for Hadoop middle tier, SAS server tier, or clients, then you must configure the cluster to map the principals correctly.

Specify the following mapping rules in hadoop.security.auth_to_local and oozie.authentication.kerberos.name.rules:

RULE: [1:$1@$0](.*@\QYOUR_AD_REALM_NAME\E$)s/@\QYOUR_AD_REALM_NAME\E$//

RULE: [2:$1@$0](.*@\QYOUR_AD_REALM_NAME\E$)s/@\QYOUR_AD_REALM_NAME\E$//

RULE: [1:$1@$0](.*@\QYOUR_MIT_REALM_NAME\E$)s/@\QYOUR_MIT_REALM_NAME\E$//

RULE: [2:$1@$0](.*@\QYOUR_MIT_REALM_NAME\E$)s/@\QYOUR_MIT_REALM_NAME\E$//

DEFAULT

*Note:* Replace `YOUR_AD_REALM_NAME` with the Active Directory realm name and `YOUR_MIT_REALM_NAME` with the Kerberos realm name.

**Additional Resources and Information**

- Go to the [SAS 9.4 Intelligence Platform Documentation page](#), and refer to the following guides for information about connecting SAS software to Kerberos environments:
  - *SAS 9.4 Intelligence Platform: Middle-Tier Administration Guide*
  - *SAS 9.4 Intelligence Platform: Security Administration Guide*

- To use a Kerberos environment, SAS Data Loader for Hadoop must be configured properly. The following tasks must be performed, typically by the SAS administrator:
  - SAS Data Loader for Hadoop must be configured for single sign on to the SAS middle tier with [Integrated Windows Authentication (IWA)](#).
  - Additional SAS configuration is required after SAS Data Loader for Hadoop is installed. For more information, see “Update SAS Configuration for Kerberos” on page 32.

- If your cluster is on a different Kerberos realm or Active Directory domain than the SAS Data Loader for Hadoop middle tier, SAS server tier, or client PC; you must configure the cluster to map the principals correctly.

- You must use a single realm for Kerberos if you plan to use SAS Data Loader for Hadoop with SAS LASR Analytic Server. SAS Data Loader for Hadoop has a Load Data to LASR directive. This directive can be used to copy Hadoop tables to a SAS LASR Analytic Server. SAS LASR Analytic Servers does not support mixed realms in a trust.

  See also “Kerberos Administrator Checklist” on page 77.

**Return to Your Scenario for the Next Step**

Select one of the following links to return to your scenario.
Deploy the SAS Embedded Process

**Before You Begin**

- The instructions in this topic are for deploying a specific release of the SAS Embedded Process on the Hadoop cluster.
  - If you have a file named `en_sasexe.zip`, then you can use the instructions in this topic.
  - If you have a file named `SAS_Viya_deployment_data.zip`, then you must use the instructions in the SAS Embedded Process: Deployment Guide to deploy, configure, upgrade, and uninstall the SAS Embedded Process on the Hadoop cluster. This release of the SAS Embedded Process is delivered from a SAS Viya repository, so the ZIP file name references SAS Viya.

  For more information about the different releases, see [information about the SAS Embedded Process on page 72](#) in the Hadoop administrator checklist.
  - Refer to the “Hadoop Administrator Checklist” on page 72.
  - You can perform one of the following methods to deploy the SAS Embedded Process:
    - Use Cloudera Manager to deploy a parcel
    - Use Ambari to deploy a stack
    - Use the SASEP-ADMIN.SH script
  - If you are following a deployment scenario, make sure that you have performed the steps in the scenario prior to this one.

**Using Cloudera Manager**

1. Make sure that you have the following file:
   ```
   SAS-Software-Depot\standalone_installs\SAS_Core_Embedded_Process_Package_for_Hadoop\13_0\Hadoop_on_Linux_x64\en_sasexe.zip
   ```
   *Note:* Ask the SAS administrator for the location of this ZIP file.

2. Unzip the `en_sasexe.zip` file to a location that you can access from the Cloudera Manager host. Here is an example:
unzip en_sasexe.zip -d sasep

3. Create a directory under the tmp directory of the file system of the host on which Cloudera Manager is installed. Here is an example: /tmp/sasep.

4. Copy the parcel directory (from the ZIP file) to the directory that you created in the previous step. Here is an example:
   cp -r sasep/sasexe/stack /tmp/sasep

5. Perform the following steps to run the install_parcel.sh script from the directory that you created.
   a. Add Execute permission to the script. Here is an example:
      ```bash
      chmod +x install_parcel.sh
      ```
   b. Run the script with sudo or root access.
      ```bash
      ./install_parcel.sh -v distro
      ```
      
      distro represents one of the following Linux distributions: redhat5, redhat6, suse11x, ubuntu10, ubuntu12, ubuntu14, debian6, or debian7. Select the appropriate value. An example is:
      ```bash
      ./install_parcel.sh -v redhat6
      ```

6. When prompted to restart Cloudera Manager, select y.
   
   **TIP**  Make sure that Cloudera Manager has restarted. If you do not restart now you can use a command such as the following to restart at a later time: service cloudera-scm-server restart.

7. Log on to Cloudera Manager.

8. Activate the SASEP parcel:
   a. From the Menu bar, select Hosts ➔ Parcels.
      
      **Note:** If the SASEP parcel is missing, run Check for new parcel.
   b. On the row for the SASEP parcel, click Distribute to copy the parcel to all nodes.
   c. Click Activate.
      
      **CAUTION:**
      
      Do not restart the cluster.
      
      After you complete the activation, you might be prompted to either restart the cluster or to close the window.
   d. If prompted, click Close.

9. Add the SASEP service to create the SASEP configuration file in HDFS.
   a. Navigate to Cloudera Manager Home.
   b. In Cloudera Manager, select the drop-down arrow next to the name of the cluster, and then select Add a Service. The Add Service Wizard appears.
   c. Select the SASEP service, and then click Continue.
   d. On the Add Service Wizard ➔ Select the set of dependencies for your new service page, select the dependencies for the service. Click Continue.
      
      **Note:** If the dependencies are automatically selected for this service, you might not see this page.
e. On the Add Service Wizard ➔ Customize Role Assignments page, select a node for the service.
   - Choose any single node.
   - **Kerberos:** Select a node that has a Kerberos keytab for the hdfs user on that node.

   Click **OK**, and then click **Continue**.

f. Enter your HDFS user name.
   - The default user name is hdfs.
   - **Kerberos:** A valid Kerberos ticket for your hdfs user must be available on the node that was selected for the SASEP service.

   **Tip** Run kinit as the hdfs user to obtain a ticket, if necessary.

g. Click **Continue**, and then click **Finish**.

h. Verify that the following file exists in the Hadoop file system: `/sas/ep/config/ep-config.xml`.

i. If the SASEP service that you just deployed is started, navigate to Cloudera Manager Home and stop the service.

10. Continue to “Return to Your Scenario for the Next Step” on page 92.

---

**Using Ambari**

1. Make sure that you have the following file:

   ```
   SAS-Software-Depot\standalone_installs\SAS_Core_Embedded_Process_Package_for_Hadoop\13_0\Hadoop_on_Linux_x64\en_sasexe.zip
   ```

   **Note:** Ask the SAS administrator for the location of this ZIP file.

2. Unzip the `en_sasexe.zip` file to a location that you can access from the Ambari host. Here is an example:

   ```
   unzip en_sasexe.zip -d sasep
   ```

3. Create a directory under the `tmp` directory of the file system of the host on which Ambari is installed. Here is an example: `/tmp/sasep`

4. Copy the `stack` directory (from the ZIP file) to the directory that you created in the previous step. Here is an example:

   ```
   cp -r sasep/sasexe/stack /tmp/sasep
   ```

5. Perform the following steps to run the `install-sasepstack.sh` script from the directory that you created.
   a. Add Execute permission to the script. Here is an example:

      ```
      chmod +x install-sasepstack.sh
      ```

   b. Run the script with sudo or root access.

      ```
      ./install_sasepstack.sh ambariAdminUsername
      ```

   c. Restart the Ambari server service, when prompted. You must restart before you can continue.
6. On the Ambari server, log on to Ambari and deploy the services:
   a. Click **Actions** and select **+ Add Service**.
      The **Add Service Wizard** page and the **Choose Services** panel appear.
   b. In the **Choose Services** panel, select the **SASEP** service. Click **Next**.
      The **Assign Slaves and Clients** panel appears.
   c. In the **Assign Slaves and Clients** panel, select the nodes where you want the stack to be deployed.
      **Note:**
      • You must select the NameNode. Select any other nodes that might run applications submitted from SAS Data Loader for Hadoop or other SAS software.
      • Depending on your version of Ambari, all nodes might be selected.
      The **Customize Services** panel appears, and the SASEP service stack is listed.
   d. Do not change any settings on the **Customize Services** panel. Click **Next**.
      **Note:** If your cluster is secured with Kerberos, the **Configure Identities** panel appears. Enter your Kerberos credentials in the **admin_principal** and **admin_password** text boxes.
      If your cluster is secured with Kerberos, the **Configure Identities** panel appears. Enter your Kerberos credentials in the **admin_principal** and **admin_password** text boxes. Click **Next**.
      The **Review** panel appears.
   e. Review the information about the panel. If everything is correct, click **Deploy**.
      The **Install, Start, and Test** panel appears. When the stack is installed on all nodes, click **Next**.
      The **Summary** panel appears.
   f. Click **Complete**. The stacks are now installed on all nodes of the cluster.
      The SASEP service is displayed on the Ambari dashboard.
   g. Verify that the following file exists in the Hadoop file system: `/sas/ep/config/ep-config.xml`.

7. Continue to “Return to Your Scenario for the Next Step” on page 92.

**Using the SASEP-ADMIN.SH Script**

**Prerequisites**
- Refer to the “Hadoop Administrator Checklist” on page 72.
- The shell script that you run to deploy the SAS Embedded Process is **sasep-admin.sh**.
- For the script to run successfully, the NameNode must connect to the DataNodes using passwordless SSH.

**CAUTION:**
If you deploy the SAS Embedded Process using the script as described below, you should also use the script if you upgrade or redeploy at a later time. Otherwise, the SAS Embedded Process can become unusable.

**Copy the Script to the Hadoop Cluster**

The sasep-admin.sh script is contained in a self-extracting archive file named sepcorehadp-13.00000-1.sh. This file is contained in a ZIP file that is put in a directory in your SAS Software Depot.

1. Create a new directory on the Hadoop NameNode that is not part of an existing directory structure, such as `/sasep`.

   This path is created on each node in the Hadoop cluster during the SAS Embedded Process installation. SAS recommends that you do not use existing system directories such as `/opt` or `/usr`. This new directory is referred to as `EPInstallDir` throughout this section.


3. Locate the `en_sasexe.zip` file. The `en_sasexe.zip` file is located in the following directory: `SAS-Software-Depot/standalone_installs/SAS_Core_Embedded_Process_Package_for_Hadoop/13_0/Hadoop_on_Linux_x64/`.

   The `sepcorehadp-13.00000-1.sh` file is included in this ZIP file.

4. Log on to the cluster using SSH.

   `ssh username@serverhostname`

5. Using a method of your choice, transfer the ZIP file to the `EPInstallDir` on your Hadoop master node. Here is an example that uses secure copy.

   `scp en_sasexe.zip username@hdpclus1:/EPInstallDir`

   **Note:** The location where you transfer the `en_sasexe.zip` file becomes the SAS Embedded Process home and is referred to as `EPInstallDir` throughout this section.

**CAUTION:**

After installation, do not delete your SAS Software Depot `standalone_installs` directory or any of its subdirectories. If hot fixes are made available for your software, they are moved to a subdirectory of the `/standalone_installs/SAS_Core_Embedded_Process_Package_for_Hadoop/` directory. The SAS Deployment Manager requires that both the initial installation files and the hot fix file exist in a subdirectory of the original SAS Software Depot `/standalone_installs/SAS_Core_Embedded_Process_Package_for_Hadoop/` directory.

**Run the Script**

To install the SAS Embedded Process and SAS Hadoop Embedded Process JAR file, follow these steps:

1. Navigate to the location on your Hadoop master node where you copied the `sepcorehadp-13.00000-1.sh` file.

   `cd /EPInstallDir`

2. Ensure that both the `EPInstallDir` folder and the `sepcorehadp-13.00000-1.sh` file have Read, Write, and Execute permissions (`chmod 755 -R`).
3. Use the following command to unpack the sepcorehadp-13.00000-1.sh file.

```bash
./sepcorehadp-13.00000-1.sh --verbose
```

**Note:** The `--quiet` option is enabled by default. Only error messages are displayed. The `--verbose` option causes all messages to be displayed that are generated during the installation process. Using verbose messaging can increase the time that is required to perform the installation.

After this script is run and the files are unpacked, the following directory structure is created where `EPInstallDir` is the location on the master node from Step 2.

- `EPInstallDir/sasexe/SASEPHome`
- `EPInstallDir/sasexe/sepcorehadp-13.00000-1.sh`

**Note:** During the install process, the `sepcorehadp-13.00000-1.sh` file is copied to all data nodes. Do not remove or move this file from the `EPInstallDir/sasexe` directory.

The SASEPHome directory should have the following structure:

- `EPInstallDir/sasexe/SASEPHome/bin`
- `EPInstallDir/sasexe/SASEPHome/install`
- `EPInstallDir/sasexe/SASEPHome/jars`
- `EPInstallDir/sasexe/SASEPHome/misc`
- `EPInstallDir/sasexe/SASEPHome/sasexe`
- `EPInstallDir/sasexe/SASEPHome/utilities`

The `EPInstallDir/SASEPHome/jars` directory contains the SAS Hadoop Embedded Process JAR file.

- `EPInstallDir/sasexe/SASEPHome/jars/sas.hadp2.jar`

The `EPInstallDir/sasexe/SASEPHome/bin` directory should contain the following script.

- `EPInstallDir/sasexe/SASEPHome/bin/sasep-admin.sh`

4. If your Hadoop cluster is secured with Kerberos and you have sudo access, the HDFS user must have a valid Kerberos ticket in order to access HDFS. You can obtain a valid Kerberos ticket with the `kinit` command.

```bash
sudo su - root
su - hdfs | hdfs-userid
kinit -kt location-of-keytab-file-user-for-which-you-are-requesting-a-ticket principal-name
exit
```

**Note:** For all Hadoop distributions except MapR, the default HDFS user is `hdfs`. For MapR distributions, the default HDFS user is `mapr`. You can specify a different user ID with the `-hdfsuser` argument when you run the `sasep-admin.sh -add` script. If you use a different HDFS superuser, ensure that the user has a home directory in HDFS before you run the `sasep-admin.sh -add` command. For example, if the HDFS superuser is `prodhdfs`, ensure that the `/user/prodhdfs` directory exists in HDFS.

**TIP** To check the status of your Kerberos ticket on the server, as the HDFS user, run the `klist` command. Here is an example of the command and its output:

```bash
klist
Ticket cache: FILE/tmp/krb5cc_493
Default principal: hdfs@HOST.COMPANY.COM

Valid starting Expires Service principal
```
5. Run the sasep-admin.sh script to deploy the SAS Embedded Process across all nodes. How you run the script depends on whether you have sudo access.

Note: It is recommended that the sasep-admin.sh script be run from the `EPInstallDir/sasexe/SASEPHome/bin/` location.

Note: Many options are available for installing the SAS Embedded Process. Review the script syntax before running it. For more information, see “Reference: SASEP-ADMIN.SH Syntax” on page 87.

• If you have sudo access, run the sasep-admin.sh script as follows to deploy SAS Embedded Process on all nodes. Review all of the information in this step and the script syntax before you run the script.

```bash
cd EPInstallDir/sasexe/SASEPHome/bin/
./sasep-admin.sh -add
```

If you have sudo access, the SAS Embedded Process install script (sasep-admin.sh) detects the Hadoop cluster topology and installs the SAS Embedded Process on all DataNode nodes. The install script also installs the SAS Embedded Process on the host node from which you run the script (the Hadoop master NameNode). The SAS Embedded Process is installed even if a DataNode is not present. To add the SAS Embedded Process to new nodes at a later time, run the sasep-admin.sh script with the `-host <hosts>` option.

In addition, a configuration file, ep-config.xml, is automatically created and written to the `EPInstallDir/SASEPHome/conf` directory and to the HDFS file system in the `/sas/ep/config` directory.

• If you do not have sudo access, follow these steps to deploy the SAS Embedded Process:

1. Run the sasep-admin.sh script as follows to deploy the SAS Embedded Process across all nodes.

```bash
cd EPInstallDir/SASEPHome/bin/
./sasep-admin.sh -x -add -hostfile host-list-filename | -host "<"host-list">
```

**CAUTION:**

The SAS Embedded Process must be installed on all nodes that are capable of running a MapReduce task (MapReduce 1) or on all nodes that are capable of running a YARN container (MapReduce 2). The SAS Embedded Process must also be installed on the host node from which you run the script (the Hadoop master NameNode). Hive and HCatalog must be available on all nodes where the SAS Embedded Process is installed. Otherwise, the SAS Embedded Process does not function properly.

**Note:** If you do not have sudo access, you must use the `-x` option and specify the hosts on which the SAS Embedded Process is deployed with either the `-hostfile` or `-host` option. Automatic detection of the Hadoop cluster topology is not available when you run the installation script with the `-x` option.

The sepcorehdp-13.00000-1.sh file is copied to all nodes that you specify. The configuration file, ep-config.xml, is created and written to the `EPInstallDir/SASEPHome/conf` directory.

2. Manually copy the ep-config.xml configuration file to the HDFS.

**Note:** This step must be performed by a user that has Write permission to the HDFS `root/` folder. If your Hadoop cluster is secured with Kerberos, the
user who copies the configuration file to HDFS must have a valid Kerberos ticket.

i. Log on as your HDFS user or as the user that you use to access HDFS.

ii. Create the `/sas/ep/config` directory for the configuration file.

    hadoop fs -mkdir -p /sas/ep/config

iii. Navigate to the `EPInstallDir/SASEPHome/conf` directory.

    cd EPInstallDir/SASEPHome/conf

iv. Use the Hadoop copyFromLocal command to copy the ep-config.xml file to HDFS.

    hadoop fs -copyFromLocal ep-config.xml /sas/ep/config/ep-config.xml

6. Verify that the SAS Embedded Process is installed by running the `sasep-admin.sh` script with the `-check` option.

   - If you ran the `sasep-admin.sh` script with sudo access, run the following command. By default, this command verifies that the SAS Embedded Process was installed on all nodes.

     cd EPInstallDir/sasexe/SASEPHome/bin/
     ./sasep-admin.sh -check

   - If you ran the `sasep-admin.sh` script with the `-x` argument, run the following command. This command verifies that the SAS Embedded Process was installed on the hosts that you specified.

     cd EPInstallDir/sasexe/SASEPHome/bin/
     ./sasep-admin.sh ./sasep-admin.sh -x -check
     -hostfile host-list-filename | -host "*"host-list"

   *Note:* The `sasep-admin.sh -check` script does not run successfully if the SAS Embedded Process is not installed.

7. Verify that the configuration file, `ep-config.xml`, was written to the HDFS file system.

    hadoop fs -ls /sas/ep/config/
    hadoop fs -cat /sas/ep/config/ep-config.xml

   *Note:* If your cluster is secured with Kerberos, you need a valid Kerberos ticket in order to access HDFS. Otherwise, you can use the WebHDFS browser.

   *Note:* The `/sas/ep/config` directory is created automatically when you run the install script with sudo access. If you used the `-genconfig` option to specify a non-default location, use that location to find the `ep-config.xml` file. When using a non-default location, a configuration property must be added to the `mapred-site.xml` configuration file that is used on the client side.

```
<property>
  <name>sas.ep.config.file</name>
  <value>config-file-location-on-hdfs</value>
</property>
```

The `config-file-location-on-hdfs` is the location of the SAS Embedded Process configuration file on HDFS.

*Reference: SASEP-ADMIN.SH Overview*

The `sasep-admin.sh` script enables you to perform the following actions.
Install or uninstall the SAS Embedded Process and SAS Hadoop MapReduce JAR file on a single node or a group of nodes.

Generate a SAS Embedded Process configuration file and write the file to an HDFS location.

Install a hot fix to the SAS Embedded Process.

Create a SAS Hadoop MapReduce JAR file symbolic link in the `hadoop/lib` directory (for backward compatibility only).

Check whether the SAS Embedded Process is installed correctly.

Display all live data nodes on the cluster.

Display the Hadoop configuration environment.

Display the Hadoop version information for the Hadoop cluster.

Display the version of the SAS Embedded Process that is installed.

The installation of the SAS Embedded Process for Hadoop involves writing a configuration file to HDFS and deploying files on all data nodes. These two tasks can occur automatically, depending on your Hadoop and HDFS permissions.

If you run the SAS Embedded Process install script (`sasep-admin.sh`) with `sudo` access, the script detects the Hadoop cluster topology and installs the SAS Embedded Process on all DataNode nodes. The install script also installs the SAS Embedded Process on the host node on which you run the script (the Hadoop master NameNode). In addition, a configuration file, `ep-config.xml`, is created and written to the HDFS file system.

If you do not have `sudo` access, you must specify the hosts on which the SAS Embedded Process is installed. In addition, you must manually copy the `ep-config.xml` configuration file to the HDFS file system.

*Note:* The Hadoop master node must be able to connect to the Hadoop slave nodes using passwordless SSH on the cluster where the SAS Embedded Process is installed.

**Reference: SASEP-ADMIN.SH Syntax**

**Action Options syntax:**

`sasep-admin.sh`

```bash
<x> -add <link> <epconfig config-filename> <maxscp number-of-copies>
<hostfile host-list-filename | -host ""host-list""""> <hdfsuser user-id>
```

`sasep-admin.sh`

```bash
<x> -genconfig <HDFS-filename> <force>
```

`sasep-admin.sh`

```bash
<x> -hotfix hotfix-filename <hostfile host-list-filename | -host ""host-list""""> <hdfsuser user-id> <maxscp number-of-copies>
```

`sasep-admin.sh`

```bash
<x> -remove <epconfig config-filename> <hostfile host-list-filename
| -host ""host-list""""> <hdfsuser user-id>
```

`sasep-admin.sh`

```bash
<x> -linklib | -unlinklib
```

**Informational options syntax:**

`sasep-admin.sh`
sasep-admin.sh

Action Arguments

-add installs the SAS Embedded Process.

Requirement If you have sudo access, the script automatically retrieves the list of data nodes from the Hadoop configuration. If you do not have sudo access, you must use the -x argument and either the -hostfile or -host argument.

Tip If you add nodes to the cluster, you can specify the hosts on which you want to install the SAS Embedded Process by using the -hostfile or -host option. The -hostfile or -host options are mutually exclusive.

See -hostfile and -host option on page 91

-genconfig <HDFS-filename> <-force>
generates a new SAS Embedded Process configuration file in the EPInstallDir/SASEPHome/conf directory of the local file system.

Requirement If you do not have sudo access, you must use the -x argument.

Interactions When used without the -x argument, the script creates the ep-config.xml configuration file and writes it to both the EPInstallDir/SASEPHome/conf directory on the local file system and the /sas/ep/config/ directory on HDFS. You can change the filename and HDFS location by using the HDFS-filename argument. HDFS-filename must be the fully qualified HDFS pathname where the configuration file is located.

When used with the -x argument, the script does not write the configuration file to HDFS. You must manually copy the file to HDFS.

Note The -genconfig argument creates two identical configuration files under EPInstallDir/SASEPHome/conf on the local file system: ep-config.xml and sasep-site.xml. The sasep-site.xml file might be copied to the client side under a folder that is in the classpath. When the sasep-site.xml file is loaded from the classpath, the configuration file on the HDFS location is not used. However, if sasep-site.xml is not found in the classpath, a configuration file must exist on HDFS, either on the default HDFS location /sas/ep/config/ep-config.xml or in the location that is set in the sas.ep.config.file property.
Tips

Use the -genconfig argument to generate a new SAS Embedded Process configuration file if you upgrade your Hadoop installation, you install or upgrade your Hive or HCatalog services, or you upgrade the JDK or JRE that is used by the Hadoop processes.

This argument generates an updated ep-config.xml file. Use the -force argument to overwrite the existing configuration file.

Use the HDFS-filename argument to specify another location and configuration filename. If you decide to generate the configuration file in a non-default HDFS location, you must set the sas.ep.config.file property in the mapred-site.xml file to the value that you specify in the -genconfig option.

See

"-epconfig config-filename" on page 91

-hotfix

-distributes a hot fix package.

Requirements

Hot fixes must be installed using the same user ID who performed the initial software installation.

Hot fixes should be installed following the installation instructions provided by SAS Technical Support.

-remove

-removes the SAS Embedded Process.

CAUTION:
Remove the QKB and the SAS Data Loader for Hadoop Spark Engine from the Hadoop nodes before removing the SAS Embedded Process.

Requirement

If you do not have sudo access, you must use the -x argument and either the -hostfile or -host argument. The -hostfile or -host options are mutually exclusive.

Interactions

When used without the -x argument and you have sudo access, the script automatically retrieves the list of data nodes from the Hadoop configuration. In addition, the script automatically removes the epconfig.xml file from HDFS.

When used with the -x argument, the SAS Embedded Process is removed from all hosts that you specify. However, the ep-config.xml file must be removed manually from HDFS.

Note

The -remove argument does not remove the SASEPHome directories. Run this command to remove the SASEPHome directories from the master node.

rm -rf SASEPHome

See

-hostfile and -host option on page 91

-linklib

-creates SAS Hadoop MapReduce JAR file symbolic links in the hadoop/lib folder.
<table>
<thead>
<tr>
<th>Restriction</th>
<th>This argument should be used only for backward compatibility (that is, when you install the July 2015 release of SAS 9.4 of the SAS Embedded Process on a client that runs SAS 9.4M2).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requirement</td>
<td>If you use this argument, you must restart the MapReduce service, the YARN service, or both after the SAS Embedded Process is installed.</td>
</tr>
<tr>
<td>Interaction</td>
<td>Use the -linklib argument after the SAS Embedded Process is already installed to create the symbolic links. Use the -link argument in conjunction with the -add argument to force the creation of the symbolic links.</td>
</tr>
<tr>
<td>See</td>
<td>“-link” on page 91</td>
</tr>
</tbody>
</table>

**-unlinklib**

removes SAS Hadoop MapReduce JAR file symbolic links in the `hadoop/lib` folder.

<table>
<thead>
<tr>
<th>Restriction</th>
<th>This argument should be used only for backward compatibility (that is, when you install the July 2015 release of SAS 9.4 of the SAS Embedded Process on a client that runs SAS 9.4M2).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requirement</td>
<td>If you use this argument, you must restart the MapReduce service, the YARN service, or both after the SAS Embedded Process is installed.</td>
</tr>
</tbody>
</table>

**Informational Arguments**

**-check**

checks whether the SAS Embedded Process is installed correctly on all data nodes.

<table>
<thead>
<tr>
<th>Requirement</th>
<th>If you ran the sasep-admin.sh script with the -x argument, you must specify the hosts for which you want to check the SAS Embedded Process by using the -hostfile or -host option. The -hostfile or -host options are mutually exclusive.</th>
</tr>
</thead>
<tbody>
<tr>
<td>See</td>
<td>-hostfile and -host option on page 91</td>
</tr>
</tbody>
</table>

**-env**

displays the SAS Embedded Process install script and the Hadoop configuration environment.

**-hadoopversion**

displays the Hadoop version information for the cluster.

**-nodefile**

displays all live DataNodes on the cluster.

| Requirement | sudo access is required. |

**-version**

displays the version of the SAS Embedded Process that is installed.

**Parameters for Action and Informational Arguments**

**-x**

if you do not have sudo access, runs the script solely under the current user’s credential.

| Requirements | This option must be the first argument passed to the script. |
A list of hosts must be provided with either the `-hostfile` or `-host` argument.

If you do not have sudo access, you must use the `-x` argument.

### Interaction

If you use the `-x` argument to install the SAS Embedded Process, that is, with the `-add` argument, you must use the `-x` argument in any other `sasep-admin.sh` script action that supports it.

### See

- `-hostfile` and `-host` option on page 91

---

**-link**

forces the creation of SAS Hadoop MapReduce JAR files symbolic links in the `hadoop/lib` folder during the installation of the SAS Embedded Process.

**Restriction**

This argument should be used only for backward compatibility (that is, when you install the July 2015 release of SAS 9.4 of the SAS Embedded Process on a client that runs SAS 9.4M2).

**Requirement**

If you use this argument, you must restart the MapReduce service, the YARN service, or both after the SAS Embedded Process is installed.

**Interactions**

Use this argument in conjunction with the `-add` argument to force the creation of the symbolic links.

Use the `-linklib` argument after the SAS Embedded Process is already installed to create the symbolic links.

**See**

- `-linklib` on page 89

---

**-epconfig** `config-filename`

generates the SAS Embedded Process configuration file in the specified location.

**Default**

If the `-epconfig` argument is not specified, the install script creates the SAS Embedded Process configuration file in the default location `/sas/ep/config/ep-config.xml`.

**Requirement**

If the `-epconfig` argument is not specified a configuration file location must be provided. If you choose a non-default location, you must set the `sas.ep.config.file` property in the `mapred-site.xml` file that is on your client machine to the non-default location.

**Interaction**

Use the `-epconfig` argument in conjunction with the `-add` or `-remove` argument to specify the HDFS location of the configuration file.

---

**-maxscp** `number-of-copies`

specifies the maximum number of parallel copies between the master and data nodes.

**Default**

10

**Interaction**

Use this argument in conjunction with the `-add` or `-hotfix` argument.

---

**-hostfile** `host-list-filename`

specifies the full path of a file that contains the list of hosts where the SAS Embedded Process is installed or removed.
Requirement

The -hostfile or -host argument is required if you do not have sudo access.

Interaction

Use the -hostfile argument in conjunction with the -add, -hotfix, -check, or -remove arguments.

See

“-hdfsuser user-id” on page 92

Example

-hostfile

-host <"> host-list <">

specifies the target host or host list where the SAS Embedded Process is installed or removed.

Requirements

If you specify more than one host, the hosts must be enclosed in double quotation marks and separated by spaces or commas.

The -host or -hostfile argument is required if you do not have sudo access.

Interaction

Use the -hostfile argument in conjunction with the -add, -hotfix, -check, or -remove arguments.

See

“-hdfsuser user-id” on page 92

Example

-host "server1 server2 server3"
-host bluesvr
-host "blue1, blue2, blue2"

-hdfsuser user-id

specifies the user ID that has Write access to HDFS root directory.

Defaults

hdfs for Cloudera, Hortonworks, Pivotal HD, and IBM BigInsights
mapr for MapR

Interactions

This argument has no affect if you use the -x argument.

Use the -hdfsuser argument in conjunction with the -add, -check, or -remove argument to change, check, or remove the HDFS user ID.

Note

The user ID is used to copy the SAS Embedded Process configuration files to HDFS.

Return to Your Scenario for the Next Step

Select one of the following links to return to your scenario.

“Scenario 1: First-time Deployment of All SAS Software” on page 6

“Scenario 2: Deployment into an Existing SAS Environment That Uses Hadoop” on page 7

“Scenario 3: Redeployment After a Hadoop Update” on page 10

“Scenario 4: Switch to a Different Hadoop Distribution” on page 11
Post-Installation Tasks for the SAS Embedded Process

Before You Begin

The post-installation tasks for the SAS Embedded Process differ by release.

• For more information, see the details about the different SAS Embedded Process releases on page 72 in the Hadoop administrator checklist.

• If the instructions in this guide were used to deploy the SAS Embedded Process, then perform the post-installation tasks in this topic.

• If the instructions in the SAS Embedded Process: Deployment Guide were used to deploy the SAS Embedded Process, then perform the post-installation tasks instructions in that guide, and do not perform the tasks in this topic.

Additional Configuration Needed to Use HCatalog File Formats

Overview of HCatalog File Types

HCatalog is a table management layer that presents a relational view of data in the HDFS to applications within the Hadoop ecosystem. With HCatalog, data structures that are registered in the Hive metastore, including SAS data, can be accessed through Hadoop code. HCatalog is part of Apache Hive.

The SAS Embedded Process for Hadoop uses HCatalog to process complex, non-delimited file formats.

Prerequisites for HCatalog Support

If you plan to access complex, non-delimited file types such as Avro or Parquet, the following conditions must be met:

• Hive must be installed on all nodes of the cluster.

• HCatalog support depends on the version of Hive that is running on your Hadoop distribution.

• For MapR distributions, Hive 0.13.0 build: 1501 or later must be installed for access to any HCatalog file type.

Setting the HCatalog CLASSPATH

If your distribution is running MapReduce 2 and YARN, the SAS Embedded Process installation automatically sets the HCatalog CLASSPATH in the ep-config.xml file. Otherwise, you must manually include the HCatalog JAR files in either the MapReduce 2 library or the Hadoop CLASSPATH.

Here is an example for a Cloudera distribution.

<property>
MapR: Adding the YARN Application CLASSPATH to the Configuration File

On all of the nodes in your Hadoop cluster that have a yarn-site.xml file, two main configuration properties specify the application CLASSPATH: yarn.application.classpath and mapreduce.application.classpath. If you do not specify the YARN application CLASSPATH, MapR takes the default CLASSPATH. However, if you specify the MapReduce application CLASSPATH, the YARN application CLASSPATH is ignored. The SAS Embedded Process for Hadoop requires both the MapReduce application CLASSPATH and the YARN application CLASSPATH.

To ensure the existence of the YARN application CLASSPATH, you must manually add the YARN application CLASSPATH to the yarn-site.xml file. Without the manual definition in the configuration file, the MapReduce application master fails to start a container.

The default YARN application CLASSPATH for Linux is:

$HADOOP_CONF_DIR,
$HADOOP_COMMON_HOME/share/hadoop/common/*,
$HADOOP_COMMON_HOME/share/hadoop/common/lib/*,
$HADOOP_HDFS_HOME/share/hadoop/hdfs/*,
$HADOOP_HDFS_HOME/share/hadoop/hdfs/lib/*,
$HADOOP_YARN_HOME/share/hadoop/yarn/*,
$HADOOP_YARN_HOME/share/hadoop/yarn/lib/*

The default YARN application CLASSPATH for Windows is:

%HADOOP_CONF_DIR%,
%HADOOP_COMMON_HOME%/share/hadoop/common/*,
%HADOOP_COMMON_HOME%/share/hadoop/common/lib/*,
%HADOOP_HDFS_HOME%/share/hadoop/hdfs/*,
%HADOOP_HDFS_HOME%/share/hadoop/hdfs/lib/*,
%HADOOP_YARN_HOME%/share/hadoop/yarn/*,
%HADOOP_YARN_HOME%/share/hadoop/yarn/lib/*
Post-Installation Tasks for the SAS Embedded Process

Note: On MapR, the YARN application CLASSPATH does not resolve the symbols or variables specified in the paths ($HADOOP_HDFS_HOME, and so on).

**Tip** Anytime you make a change to the cluster, remember to restart the node managers to pick up the change.

(Optional) Changing the Trace Level

You can modify the level of tracing by changing the value of the sas.ep.server.trace.level property in the ep-config.xml file. The default value is 4 (TRACE_NOTE).

```xml
<property>
  <name>sas.ep.server.trace.level</name>
  <value>trace-level</value>
</property>
```

The `trace-level` represents the level of trace that is produced by the SAS Embedded Process. `trace-level` can be one of the following values:

- 0 TRACE_OFF
- 1 TRACE_FATAL
- 2 TRACE_ERROR
- 3 TRACE_WARN
- 4 TRACE_NOTE
- 5 TRACE_INFO
- 10 TRACE_ALL

*Note:* When using SAS Embedded Process with SAS High-Performance Analytics (HPA), tracing requires that an `/opt/SAS` directory exist on every node of the cluster where the SAS Embedded Process is installed. If the folder does not exist or does not have Write permission, the SAS Embedded Process job fails.

(Optional) Adjusting the SAS Embedded Process Performance

You can adjust how the SAS Embedded Process runs by changing or adding its properties. SAS Embedded Process configuration properties can be added to the mapred-site.xml configuration on the client side or the sasep-site.xml file. If you change the properties in the ep-config.xml file located on HDFS, it will affect all SAS Embedded Process jobs. The ep-config.xml file is created when you install the SAS Embedded Process.

When using the SAS Embedded Process, there are a number of properties that you can adjust to improve performance.

- You can specify the number of SAS Embedded Process MapReduce tasks per node by changing the sas.ep.superreader.tasks.per.node property.
The SAS Embedded Process super reader technology does not use the standard MapReduce split calculation. Instead of assigning one split per task, it assigns many. The super reader calculates the splits, groups them, and distributes the groups to a configurable number of mapper tasks based on data locality.

The default number of tasks is 6.

```
<property>
  <name>sas.ep.superreader.tasks.per.node</name>
  <value>number-of-tasks</value>
</property>
```

- You can specify the number of concurrent nodes that are allowed to run High-Performance Analytics output tasks by changing the `sas.ep.hpa.output.concurrent.nodes` property.

  If this property is set to zero, the SAS Embedded Process allocates tasks on all nodes capable of running a YARN container. If this property is set to −1, the number of concurrent nodes equates to the number of High-Performance Analytics worker nodes. If the number of concurrent nodes exceeds the number of available nodes, the property value is adjusted to the number of available nodes. The default value is 0.

```
<property>
  <name>sas.ep.hpa.output.concurrent.nodes</name>
  <value>number-of-nodes</value>
</property>
```

- You can specify the number of High-Performance Analytics output tasks that are allowed to run per node by changing the `sas.ep.hpa.output.tasks.per.node` property.

  The default number of tasks is 1.

```
<property>
  <name>sas.ep.hpa.output.tasks.per.node</name>
  <value>number-of-tasks</value>
</property>
```

- You can specify the number of concurrent input reader threads.

  Each reader thread takes a file split from the input splits queue, opens the file, positions itself at the beginning of the split, and starts reading the records. Each record is stored on a native buffer that is shared with the DS2 container. When the native buffer is full, it is pushed to the DS2 container for processing. When a reader thread finishes reading a file split, it takes another file split from the input splits queue. The default number of input threads is 3.

```
<property>
  <name>sas.ep.input.threads</name>
  <value>number-of-input-threads</value>
</property>
```

- You can specify the number of output writer threads by changing the `sas.ep.output.threads` property.

  The SAS Embedded Process super writer technology improves performance by writing output data in parallel, producing multiple parts of the output file per mapper task. Each writer thread is responsible for writing one part of the output file. The default number of output threads is 2.

```
<property>
  <name>sas.ep.output.threads</name>
  <value>number-of-output-threads</value>
</property>
```
• You can specify the number of compute threads by changing the sas.ep.compute.threads property.

Each compute thread runs one instance of the DS2 program inside the SAS Embedded Process. The DS2 code that runs inside the DS2 container processes the records that it receives. At a given point, DS2 flushes output data to native buffers. The super writer threads take the output data from DS2 buffers and writes them to the super writer thread output file on a designated HDFS location. When all file input splits are processed and all output data is flushed and written to HDFS, the mapper task ends. The default number of compute threads is 1.

```
<property>
  <name>sas.ep.compute.threads</name>
  <value>number-of-threads</value>
</property>
```

• You can specify the number of buffers that are used for output data by changing the sas.ep.output.buffers property.

The number of output buffers should not be less than sas.ep.compute.threads plus sas.ep.output.threads. The default number of buffers is 3.

```
<property>
  <name>sas.ep.output.buffers</name>
  <value>number-of-buffers</value>
</property>
```

• You can specify the number of native buffers that are used to cache input data by changing the sas.ep.input.buffers property. The default value is 4. The number of input buffers should not be less than sas.ep.compute.threads plus sas.ep.input.threads.

```
<property>
  <name>sas.ep.input.buffers</name>
  <value>number-of-buffers</value>
</property>
```

• You can specify the optimal size of one input buffer by changing the sas.ep.optimal.input.buffer.size property.

The optimal row array size is calculated based on the optimal buffer size. The default value is 1 MB.

```
<property>
  <name>sas.ep.optimal.input.buffer.size.mb</name>
  <value>number-in-megabytes</value>
</property>
```

• You can specify the amount of memory in bytes that the SAS Embedded Process is allowed to use with MapReduce 1 by changing the sas.ep.max.memory property in the ep-config.xml file.

Adjust the YARN container limit to change the amount of memory that the SAS Embedded Process is allowed to use.

The default value is 2147483647 bytes.

```
<property>
  <name>sas.ep.max.memory</name>
  <value>number-of-bytes</value>
</property>
```
Deploy the SAS Data Loader for Hadoop Spark Engine

Before You Begin

- Refer to the “Hadoop Administrator Checklist” on page 72.
- You can perform one of the following methods to deploy the SAS Data Loader for Hadoop Spark Engine:
  - Use Cloudera Manager to deploy a parcel
  - Use Ambari to deploy a stack
  - Use the SASDMP_ADMIN.SH Script
- If you are following a deployment scenario, make sure that you have performed the steps in the scenario prior to this one.

Using Cloudera Manager

1. Make sure that you have the following file:
   ```
   SAS-Software-Depot\standalone_installs\SAS_Data_Loader_for_Hadoop_Spark_Engine\3_1\Hadoop_on_Linux_x64\en_sasexe.zip
   ```
   
   *Note:* Ask the SAS administrator for the location of this ZIP file.

2. Unzip the `en_sasexe.zip` file to a location that you can access from the Cloudera Manager host. Here is an example:
   ```
   unzip en_sasexe.zip -d sasspark
   ```

3. Create a directory under the `tmp` directory of the file system of the host on which Cloudera Manager is installed. Here is an example: `/tmp/sasspark.`
4. Copy the parcel directory (from the ZIP file) to the directory that you created in the previous step. Here is an example:
   
   ```bash
   cp -r saasspark/sasexe/parcel /tmp/saasspark
   ```

5. Perform the following steps to run the `install_parcel.sh` script from the directory that you created.
   a. Add Execute permission to the script. Here is an example:
      
      ```bash
      chmod +x install_parcel.sh
      ```
   b. Run the script with sudo or root access.
      
      ```bash
      ./install_parcel.sh -v distro
      ```

      `distro` represents one of the following Linux distributions: redhat5, redhat6, suse11x, ubuntu10, ubuntu12, ubuntu14, debian6, or debian7. Select the appropriate value. An example is:
      
      ```bash
      ./install_parcel.sh -v redhat6
      ```

6. When prompted to restart Cloudera Manager, select y.
   
   **TIP** Make sure that Cloudera Manager has restarted. If you do not restart now you can use a command such as the following to restart at a later time: `service cloudera-scm-server restart`.

7. Log on to Cloudera Manager.

8. Activate the SASDMSPARK parcel:
   a. From the Menu bar, select **Hosts** → **Parcels**.
      
      **Note:** If the SASDMSPARK parcel is missing, run **Check for new parcel**.
   b. On the row for the SASDMSPARK parcel, click **Distribute** to copy the parcel to all nodes.
   c. Click **Activate**.
      
      **CAUTION:**
      
      Do not restart the cluster.
      
      After you complete the activation, you might be prompted to either restart the cluster or to close the window.
   d. If prompted, click **Close**.

9. Add the SASDMSPARK service to create the SASDMSPARK configuration file in HDFS.
   a. Navigate to Cloudera Manager Home.
   b. In Cloudera Manager, select the drop-down arrow next to the name of the cluster, and then select **Add a Service**. The Add Service Wizard appears.
   c. Select the SASDMSPARK service, and then click **Continue**.
   d. On the **Add Service Wizard** → **Select the set of dependencies for your new service** page, select the dependencies for the service. Click **Continue**.
      
      **Note:** If the dependencies are automatically selected for this service, you might not see this page.
   e. On the **Add Service Wizard** → **Customize Role Assignments** page, select a node for the service.
      
      • Choose any single node.
• **Kerberos:** Select a node that has a Kerberos keytab for the hdfs user on that node.

Click **OK**, and then click **Continue**.

f. If prompted, enter your HDFS user name.

• The default user name is hdfs.

• **Kerberos:** A valid Kerberos ticket for your hdfs user name must be available on the node that was selected for the SASDMSpark service.

  **Tip**  Run `kinit` as the hdfs user to obtain a ticket, if necessary.

g. Click **Continue**, and then click **Finish**.

h. Verify that the following file exists in the Hadoop file system: `/sas/ep/config/dmp-config.xml`.

i. If the SASDMSpark service that you just deployed is started, navigate to Cloudera Manager Home and stop the service.

10. Continue to “Return to Your Scenario for the Next Step” on page 107.

**Using Ambari**

1. Make sure that you have the following file:

   ```
   SAS-Software-Depot\standalone_installs\SAS_Data_Loader_for_Hadoop_Spark_Engine\3_1\Hadoop_on_Linux_x64\en_sasexe.zip
   ```

   **Note:** Ask the SAS administrator for the location of this ZIP file.

2. Unzip the `en_sasexe.zip` file to a location that you can access from the Ambari host.

   Here is an example:

   ```
   unzip en_sasexe.zip -d sasspark
   ```

3. Create a directory under the `tmp` directory of the file system of the host on which Ambari is installed. Here is an example: `/tmp/sasspark`

4. Copy the `stack` directory (from the ZIP file) to the directory that you created in the previous step. Here is an example:

   ```
   cp -r ep/sasexe/stack /tmp/sasspark
   ```

5. Perform the following steps to run the `install-sasdmspark.sh` script from the directory that you created.

   a. Add Execute permission to the script. Here is an example:

      ```
      chmod +x install-sasdmspark.sh
      ```

   b. Run the script with `sudo` or root access.

      ```
      ./install_sasdmspark.sh ambariAdminUsername
      ```

   c. Restart the Ambari server service, when prompted. You must restart before you can continue.

      **Note:** The script can restart the Ambari server service for you, or you can do it manually.

6. On the Ambari server, log on to Ambari and deploy the services:

   a. Click **Actions** and select + **Add Service**.
Add Service Wizard page and the Choose Services panel appear.

b. In the Choose Services panel, select the SASDMSPARK service. Click Next.

The Assign Slaves and Clients panel appears.

c. In the Assign Slaves and Clients panel, select the nodes where you want the stack to be deployed.

Note:
- You must select the NameNode. Select any other nodes that might run applications submitted from SAS Data Loader for Hadoop or other SAS software.
- Depending on your version of Ambari, all nodes might be selected.

The Customize Services panel appears, and the SASDMSPARK service stack is listed.

d. Do not change any settings on the Customize Services panel. Click Next.

Note: If your cluster is secured with Kerberos, the Configure Identities panel appears. Enter your Kerberos credentials in the admin_principal and admin_password text boxes.

If your cluster is secured with Kerberos, the Configure Identities panel appears. Enter your Kerberos credentials in the admin_principal and admin_password text boxes. Click Next.

The Review panel appears.

e. Review the information about the panel. If everything is correct, click Deploy.

The Install, Start, and Test panel appears. When the stack is installed on all nodes, click Next.

The Summary panel appears.

f. Click Complete. The stacks are now installed on all nodes of the cluster.

The SASDMSPARK service is displayed on the Ambari dashboard.

g. Verify that the following file exists in the Hadoop file system: `/sas/ep/config/dmp-config.xml`.

7. Continue to “Return to Your Scenario for the Next Step” on page 107.

Using the SASDMP_ADMIN.SH Script

Prerequisites
- Refer to the “Hadoop Administrator Checklist” on page 72.
- The script that deploys the SAS Data Loader for Hadoop Spark Engine is sasdmp_admin.sh.
- You must have sudo access on the NameNode only to run the sasdmp_admin.sh script. Also, you must have SSH set up so that the NameNode can use passwordless SSH to all DataNodes on the cluster where Data Loader for Hadoop Spark Engine is installed.

CAUTION:
- If you deploy the SAS Data Loader for Hadoop Spark Engine using the script as described below, you should also use the script if you upgrade or redeploy at a
later time. Otherwise, the SAS Data Loader for Hadoop Spark Engine can become unusable.

**Copy the Script to the Hadoop Cluster**
The SAS Data Loader for Hadoop Spark Engine install script is contained in a self-extracting archive file named sasdmp_admin.sh. This file is contained in a ZIP file that is located in a directory in your SAS Software Depot.

To copy the ZIP file to the SPRKInstallDir on your Hadoop master node, follow these steps:

1. Navigate to the **SAS-Software-Depot/standalone_installs** directory.
   
   This directory was created when your SAS Software Depot was created by the SAS Deployment Wizard.

2. Locate the en_sasexe.zip file. This file is in the following directory: **SAS-Software-Depot/standalone_installs/SAS_Data_Loader_for_Hadoop_Spark_Engine/3_1/Hadoop_on_Linux_x64**.
   
   The sasdmp_admin.sh file is included in this ZIP file.

3. Log on to the cluster using SSH.
   
   ```sh
   ssh username@serverhostname
   ```

4. Copy the en_sasexe.zip file from the client to the **SPRKInstallDir** on the cluster.

   ```sh
   Note: The **SPRKInstallDir** location becomes the SAS Data Loader for Hadoop Spark Engine home.
   ```

**Run the Script**

To install SAS Data Loader for Hadoop Spark Engine, follow these steps:

1. Navigate to the location on your Hadoop master node where you copied the en_sasexe.zip file.
   
   ```sh
   cd /SPRKInstallDir
   ```

2. Unzip the en_sasexe.zip file.

   ```sh
   unzip en_sasexe.zip
   ```

   After the file is unzipped, a **sasexe** directory is created in the same location as the en_sasexe.zip file. The dmsprkhadp-3.10000-1.sh file is located in the **sasexe** directory.

   ```sh
   SPRKInstallDir/sasexe/dmsprkhadp-3.10000-1.sh
   ```

3. Ensure that both the **SPRKInstallDir** folder and the en_sasexe.zip file have Read, Write, and Execute permissions (**chmod 755**).

4. Navigate to the **sasexe** directory.

   ```sh
   cd /sasexe
   ```

5. Use the following command to unpack the dmsprkhadp-3.10000-1.sh file.

   ```sh
   ./dmsprkhadp-3.10000-1.sh
   ```

   **Note:** The exact version of the script might vary.

   After this script is run and the files are unpacked, the script creates the following directory structure:
SPRKInstallDir/sasexe/SASDMPHome
SPRKInstallDir/sasexe/dmsprkhadp-3.10000-1.sh

Note: During the install process, the dmsprkhadp-3.10000-1.sh is copied to all data nodes. Do not remove or move this file from the SPRKInstallDir/sasexe directory.

An example of the SASDMPHome directory structure is shown below. Your directories might differ.

SPRKInstallDir/sasexe/SASDMPHome/bin
SPRKInstallDir/sasexe/SASDMPHome/dat
SPRKInstallDir/sasexe/SASDMPHome/etc
SPRKInstallDir/sasexe/SASDMPHome/lib
SPRKInstallDir/sasexe/SASDMPHome/share
SPRKInstallDir/sasexe/SASDMPHome/var

An example of SPRKInstallDir/sasexe/SASDMPHome/bin is shown below. Your files might differ.

SPRKInstallDir/sasexe/SASDMPHome/bin/dfwsvc
SPRKInstallDir/sasexe/SASDMPHome/bin/dfxver
SPRKInstallDir/sasexe/SASDMPHome/bin/dfxver.bin
SPRKInstallDir/sasexe/SASDMPHome/bin/sasdmp_admin.sh
SPRKInstallDir/sasexe/SASDMPHome/bin/settings.sh
SPRKInstallDir/sasexe/SASDMPHome/bin/dmpsvc

6. Use the sasdmp_admin.sh script to deploy the SAS Data Loader for Hadoop Spark Engine installation across all nodes.

Note: Many options are available when installing SAS Data Loader for Hadoop Spark Engine. Review the script syntax before running it. For more information, see “Reference: SASDMP_ADMIN.SH Overview” on page 104.

If your cluster is secured with Kerberos, complete both steps Step 6a on page 103 and Step 6b on page 103. If your cluster is not secured with Kerberos, complete only Step 6b on page 103.

a. If your cluster is secured with Kerberos, the HDFS user must have a valid Kerberos ticket to access HDFS. This can be done with kinit.

```
sudo su - root
su - hdfs | hdfs-userid
kinit -kt location-of-keytab-file-user-for-which-you-are-requesting-a-ticket
exit
```

Note: For all Hadoop distributions except MapR, the default HDFS user is hdfs. For MapR distributions, the default MapR superuser is mapr. You can specify a different user ID with the -hdfsuser argument when you run the bin/sasdmp_admin.sh -add script.

Tip: To check the status of your Kerberos ticket on the server, run klist while you are running as the -hdfsuser user.

b. Run the sasdmp_admin.sh script. Review all of the information in this step before running the script.

```
cd SPRKInstallDir/sasexe/SASDMPHome/bin/
./sasdmp_admin.sh -add
./sasdmp_admin.sh -genconfig
```

Note: You must install SAS Data Loader for Hadoop Spark Engine on all nodes that might run applications submitted from SAS Data Loader for Hadoop. By default,
the installation script installs SAS Data Loader for Hadoop Spark Engine on the host where you run the script, the NameNode, and all DataNode hosts. This occurs even if a DataNode is not present.

**Note:** You can run the script with the `-host <hosts>` option for the following cases:

- To add SAS Data Loader for Hadoop Spark Engine to new nodes at a later time
- To deploy the software to any node that has a YARN NodeManager

7. Verify that Data Loader for Hadoop Spark Engine is installed by running the `sasdmp_admin.sh` script with the `-check` option.

```bash
cd SPRKInstallDir/sasexe/SASDMPHome/bin/
./sasdmp_admin.sh -check
```

This command checks whether SAS Data Loader for Hadoop Spark Engine is installed on all data nodes.

**Note:** The `sasdmp_admin.sh -check` script does not run successfully if SAS Data Loader for Hadoop Spark Engine is not installed.

8. Verify that the configuration file, dmp-config.xml, was written to the HDFS file system.

```bash
hadoop fs -ls /sas/ep/config/
hadoop fs -cat /sas/ep/config/dmp-config.xml
```

**Note:** If your cluster is secured with Kerberos, you need a valid Kerberos ticket to access HDFS. If not, you can use the WebHDFS browser.

**Note:** The `/sas/ep/config` directory is created automatically when you run the install script. If you used `-dmpconfig` or `-genconfig` to specify a non-default location, use that location to find the dmp-config.xml file.

**Reference: SASDMP_ADMIN.SH Overview**

The `sasdmp_admin.sh` script enables you to perform the following actions.

- Install or uninstall SAS Data Loader for Hadoop Spark Engine on a single node or a group of nodes.
- Check if SAS Data Loader for Hadoop Spark Engine is installed correctly.
- Generate a SAS Data Loader for Hadoop Spark Engine configuration file and write the file to an HDFS location.
- Write the installation output to a log file.
- Display all live data nodes on the cluster.
- Display the Hadoop configuration environment.

**Note:** You must have sudo access on the NameNode only to run the `sasdmp_admin.sh` script. Also, You must also have SSH set up so that the NameNode can passwordless SSH to all DataNodes on the cluster where SAS Data Loader for Hadoop Spark Engine is installed.

**Reference: SASDMP_ADMIN.SH Syntax**

```bash
sasdmp_admin.sh
```

```
-add <-dmpconfig config-filename > <maxscp number-of-copies>
<-hostfile host-list-filename | -host <"host-list">>
<-hdfsuser user-id > <-log filename>
```
sasdpmp_admin.sh
- remove <-dmpconfig config-filename > <-hostfile host-list-filename | -host <"host-
lis">>
<-hdfsuser user-id> <-log filename><-keepconfig>

sasdpmp_admin.sh
<-genconfig config-filename <-force>>
<-check> <-hostfile host-list-filename | -host <"host-list">>
<-env>
<-hadoopversion >
<-hottfix >
<-log filename>
<-nodefile>
<-sparkversion>
<-validate>
<-version >

Arguments
- add
installs SAS Data Loader for Hadoop Spark Engine.

Tip To add nodes to the cluster, you can specify the hosts on which you want to
install the SAS Data Loader for Hadoop Spark Engine by using the -hostfile or
-host option. The -hostfile and -host options are mutually exclusive.

See -hostfile and -host option on page 105

-dmpconfig config-filename
generates the SAS Data Loader for Hadoop Spark Engine configuration file in the
specified location.

Default /sas/ep/config/dmp-config.xml

Interaction Use the -dmpconfig argument in conjunction with the -add or -remove
argument to specify the HDFS location of the configuration file.

Tip Use the -dmpconfig argument to create the configuration file in a non-
default location.

See “-genconfig config-filename -force” on page 107

-maxscp number-of-copies
specifies the maximum number of parallel copies between the master and data nodes.

Default 10

Interaction Use this argument in conjunction with the -add argument.

-hostfile host-list-filename
specifies the full path of a file that contains the list of hosts where SAS Data Loader
for Hadoop Spark Engine is installed or removed.

Default The sasdpmp_admin.sh script discovers the cluster topology and uses the
retrieved list of data nodes.
Interaction Use the -hostfile argument in conjunction with the -add when new nodes are added to the cluster.

Tip You can also assign a host list filename to a UNIX variable, SASEP_HOSTS_FILE.

```
export SASEP_HOSTS_FILE=/etc/hadoop/conf/slaves
```

See “-hdfsuser user-id” on page 106

Example

```
-hostfile /etc/hadoop/conf/slaves
```

### -host <""> host-list <""

specifies the target host or host list where SAS Data Loader for Hadoop Spark Engine is installed or removed.

**Default** The sasdmp_admin.sh script discovers the cluster topology and uses the retrieved list of data nodes.

**Requirement** If you specify more than one host, the hosts must be enclosed in double quotation marks and separated by spaces.

Interaction Use the -host argument in conjunction with the -add when new nodes are added to the cluster.

Tip You can also assign a list of hosts to a UNIX variable, SASEP_HOSTS.

```
export SASEP_HOSTS="server1 server2 server3"
```

See “-hdfsuser user-id” on page 106

Example

```
-host "server1 server2 server3"
-host bluesvr
```

### -hdfsuser user-id

specifies the user ID that has Write access to the HDFS root directory.

**Defaults**

- hdfs for Cloudera, Hortonworks, Pivotal HD, and IBM BigInsights
- mapr for MapR

**Interaction** Use the -hdfsuser argument in conjunction with the -add or -remove argument to change or remove the HDFS user ID.

**Note** The user ID is used to copy the Data Loader for Hadoop Spark Engine configuration files to HDFS.

### -log filename

writes the installation output to the specified filename.

**Interaction** Use the -log argument in conjunction with the -add or -remove argument to write or remove the installation output file.

### -remove <-keepconfig>

removes SAS Data Loader for Hadoop Spark Engine.
Tips
You can specify the hosts for which you want to remove SAS Data Loader for Hadoop Spark Engine by using the -hostfile or -host option. The -hostfile or -host options are mutually exclusive.

This argument removes the generated dmp-config.xml file. Use the -keepconfig argument to retain the existing configuration file.

See -hostfile and -host option on page 106

-genconfig config-filename <-force>
generates a new SAS Data Loader for Hadoop Spark Engine configuration file in the specified location.

Default /sas/ep/config/dmp-config.xml

Interaction Use the -genconfig argument in conjunction with the -add or -remove argument to specify the HDFS location of the configuration file.

Tip This argument generates an updated dmp-config.xml file. Use the -force argument to overwrite the existing configuration file.

See “-dmpconfig config-filename” on page 105

-check checks if SAS Data Loader for Hadoop Spark Engine is installed correctly on all data nodes.

-env displays the Hadoop configuration environment.

-hadoopversion displays the Hadoop version information for the cluster.

-hotfix installs a hotfix on an existing SAS Data Loader for Hadoop Spark Engine installation.

-nodelist displays all live DataNodes on the cluster.

-sparkversion displays the Spark version information for the cluster.

-validate validates the install by executing simple Spark and MapReduce jobs.

-version displays the version of SAS Data Loader for Hadoop Spark Engine that is installed.

Return to Your Scenario for the Next Step
Select one of the following links to return to your scenario.

“Scenario 1: First-time Deployment of All SAS Software” on page 6

“Scenario 2: Deployment into an Existing SAS Environment That Uses Hadoop” on page 7

“Scenario 3: Redeployment After a Hadoop Update” on page 10
Use the Hadoop Tracer Script to Collect Hadoop JAR and Configuration Files

Overview

SAS Data Loader for Hadoop requires Hadoop JAR and configuration files from the Hadoop cluster. You must collect the Hadoop JAR and configuration files for the following reasons:

- You want to deploy SAS Data Loader for Hadoop.
- You updated your Hadoop cluster after SAS Data Loader for Hadoop has been deployed. For example, the Hadoop cluster is updated with a new parcel, package, service, or component on an existing cluster; or you make any configuration changes to the Hadoop services or components.
- You switch to a new Hadoop distribution after SAS Data Loader for Hadoop has been deployed. For example, you want to switch from Cloudera to Hortonworks.

You can collect the Hadoop JAR and configuration files in one of the following ways:

- **Run the Hadoop tracer script on page 109** as described below.
- Collect the files by using tools that are specific to deploying and administering SAS software. Typically, a SAS administrator uses these tools.
  - Use the SAS Deployment Wizard to collect the files before SAS Data Loader for Hadoop is deployed.
  - Use the SAS Deployment Manager to collect the files after SAS Data Loader for Hadoop software is deployed.

*Note:* The person running the tools must have administrator credentials for the Hadoop cluster. If the SAS Deployment Wizard or SAS Deployment Manager is used, you do not need to run the Hadoop tracer script.

*Note:* If files are missing or invalid, SAS Data Loader for Hadoop will not operate successfully.

Prerequisites

SAS provides a Hadoop tracer script to collect the Hadoop JAR and configuration files. To run the Hadoop tracer script successfully:

- Ensure that the user running the script has a UNIX account with SSH, either password or private key to the Hive node or NameNode.
- Ensure that Python 2.6 or later and strace are installed. Contact your system administrator if these packages are not installed on the system.
Ensure that the user running the script has authorization to issue HDFS and Hive commands.

If Hadoop is secured with Kerberos, obtain a Kerberos ticket for the user before running the script.

Ensure that the user running the script has authorization to issue HDFS and Hive commands. One way to do so is to run the following command:

```
$ time hive -e 'set -v'
```

Run the `hadoop` and `hive` commands on the Hadoop cluster. If these commands fail, the script will not run.

---

**Run the Hadoop Tracer Script**

To obtain and run the Hadoop tracer script:

1. On the Hadoop server, create a temporary directory, such as `~/opt/sas/hadoopfiles/temp`.

2. Paste the following address in a web browser to download the hadooptracer.zip file to the directory that you created in the previous step.


3. Using a transfer method, such as PSFTP, SFTP, SCP, or FTP, transfer the ZIP file to the Hive node on your Hadoop cluster.

4. Unzip the file. The hadooptracer.py file is included in this ZIP file.

5. Change permissions on the file to have EXECUTE permission.

   `chmod 755 ./hadooptracer.py`

6. Run the following command with required options:

   `python hadooptracer.py --filterby=latest --postprocess -f /tmp/sitexmls/hadooptracer.json`

**Note:**

- The Hadoop tracer script creates JAR files and configuration files in the `/tmp/jars` and `/tmp/sitexmls` directories, respectively. The SAS administrator needs access to the files in these directories in order to deploy SAS Data Loader for Hadoop.

- The options shown are:
  - `--filterby=latest` ensures that if duplicate JAR or configuration files exist, the latest version is selected.
  - `--postprocess` removes version numbers from configuration files that are collected. This option is ignored for Cloudera clusters.
  - `-f` creates a custom path and filename for the output file.
  - The `--winplatforms` option performs post-processing required for SAS Clients that will be deployed on Windows; such as the SAS Data Loader middle tier, or the SAS Workspace Server.
  - The `--nothreads` option (not shown) is useful for small Hadoop clusters or those with low memory.
  - To see additional options, run the command with the `-h` option.

- Additional JAR and configuration files might be needed because of JAR file interdependencies and your Hadoop distributions.
• The Hadoop tracer script creates the hadooptracer.log file in the /tmp
directory. Some error messages in the console output for hadooptracer.py are
normal and do not necessarily indicate a problem with the JAR and
configuration file collection process. However, if the files are not collected as
expected or if you experience problems connecting to Hadoop with the
collected files, contact SAS Technical Support and include the
hadooptracer.log file.

7. Delete the following files, if found:
   • derby*.jar
   • spark-examples*.jar
   • ranger-plugins-audit*.jar
   • avatica*.jar
   • hadoop-0.20.2-dev-core*.jar
   • jackson-core*.jar
   • htrace-core-3.1.0-incubating.jar
   • calcite-avatica-1.6.0.jar

8. Provide the /tmp/jars and /tmp/sitexmls directories to the SAS administrator.

Return to Your Scenario for the Next Step

Select one of the following links to return to your scenario.

“Scenario 1: First-time Deployment of All SAS Software” on page 6
“Scenario 2: Deployment into an Existing SAS Environment That Uses Hadoop” on page 7
“Scenario 3: Redeployment After a Hadoop Update” on page 10
“Scenario 4: Switch to a Different Hadoop Distribution” on page 11
“Scenario 5: Upgrade Software” on page 13

Deploy the SAS Quality Knowledge Base

Before You Begin

• You can perform one of the following methods to deploy the SAS Quality
   Knowledge Base:
   • Use Cloudera Manager to deploy a parcel that is created by using SAS
     Deployment Manager
   • Use Ambari to deploy a stack that is created by using SAS Deployment Manager
   • Use the QKB_PUSH.SH script
• If you are following a deployment scenario, make sure that you have performed the steps in the scenario prior to this one.

Using Cloudera or Ambari

Prerequisites
Consider the following requirements:

• To deploy a QKB using Cloudera Manager or Ambari, you must perform the following steps:

  1. Run SAS Deployment Manager to create a parcel or stack of the QKB, and copy it to the cluster manager node on your Hadoop cluster.

     *Note:* You must know the name of the directory that contains the QKB, and typically a SAS administrator can provide it. For more information, see “Prepare the SAS Quality Knowledge Base for Deployment” on page 29.

  2. Run Cloudera Manager or Ambari to deploy the parcel or stack, respectively, to the DataNodes of the Hadoop cluster.

• The Hadoop cluster must be based on UNIX, and the SAS server tier must be based on UNIX. The task to run in the SAS Deployment Manager is available only on UNIX.

• The Hadoop distribution must be a supported version of Cloudera CDH or Hortonworks.

• If you want to replace a QKB on the Hadoop cluster, you must remove the old one first. For more information, see “Remove SAS Software from the Hadoop Cluster” on page 119.

• If you have already deployed the QKB by using a script on page 114, do not perform the instructions in this topic.

To create the parcel or stack using the SAS Deployment Manager, you must have the following information:

• the host name and port number for your Hadoop cluster manager host

• a UNIX user account with SSH for the cluster manager host

• credentials for the cluster manager application (Cloudera Manager or Ambari): this is a UNIX account name that has administrator privileges and a password for the cluster manager software

• the directory for the SAS QKB to be deployed

• instructions for logging on to the cluster manager software

To deploy the parcel or stack of the QKB by using Cloudera Manager or Ambari:

• You must have Write permission to the root of HDFS.

• If your Hadoop cluster is secured with Kerberos, use `kinit` to acquire a Kerberos ticket for the HDFS user who runs the cluster manager.
Create a Parcel or Stack of the QKB

Perform the following steps to run SAS Deployment Manager to create a parcel or stack of the QKB, and copy it to the Hadoop cluster manager node.

1. If you have not done so already, obtain a QKB as described in “Prepare the SAS Quality Knowledge Base for Deployment” on page 29.

2. Open a command prompt on a machine in the SAS server tier.

3. Navigate to the SAS Deployment Manager directory. An example directory might be: `<SASHome>/SASDeploymentManager/9.4`

4. On UNIX, enter `.sasadm.sh`.

5. Select a language and click OK. The first page of the SAS Deployment Manager is displayed.


7. Select your Hadoop distribution (Cloudera or Hortonworks) and click Next.

8. Enter the host name and port number for your Hadoop cluster manager host machine. Click Next.

9. Enter credentials for the Hadoop cluster manager application. This is a UNIX account name that has administrator privileges and a password for your cluster manager software (Cloudera Manager or Ambari). Click Next.

10. Enter credentials for the Hadoop cluster manager host. This is a UNIX user account with SSH for the Hadoop cluster manager host machine. Click Next.

11. Specify the directory where the SAS QKB to be deployed is located.

12. Accept the default temporary installation directory or specify a new one. Click Next.

13. Specify a site-specific version number for the parcel or stack (stack) to be deployed. Click Next.

14. Allow the SAS Deployment Wizard to verify that the installation can proceed. Click Next.

15. Review the status of the QKB deployment. If all is well, click Next.

16. Click through the next several pages. Then click Finish.

   The task generates a Cloudera parcel or a Hortonworks stack for the specified QKB. The task copies the parcel or stack to the cluster manager node on your Hadoop cluster.

17. Continue to one of the following tasks to deploy the parcel or stack to the DataNodes on the Hadoop cluster.

   • “Deploy the Parcel Using Cloudera Manager” on page 112
   • “Deploy the Stack Using Ambari” on page 113

Deploy the Parcel Using Cloudera Manager

1. Make sure that you have created the parcel and copied it to the Hadoop cluster manager node, as described in “Create a Parcel or Stack of the QKB” on page 112.

2. Log on to Cloudera Manager.

3. Deploy the QKB to the Hadoop cluster:

   a. From the Menu bar, select Hosts ⇒ Parcels.

      Note: If the SASQKB parcel is missing, run Check for new parcel.
b. On the row for the SASQKB parcel, click Distribute to copy the parcel to all nodes.

c. Click Activate.

**CAUTION:**
Do not restart the cluster.

After you complete the activation, you might be prompted to either restart the cluster or to close the window.

d. If prompted, click Close.

4. Add the SASQKB service to create the SASQKB index file in HDFS.

   a. Navigate to Cloudera Manager Home.
   
   b. In Cloudera Manager, select the drop-down arrow next to the name of the cluster, and then select Add a Service. The Add Service Wizard appears.
   
   c. Select the SASQKB service, and then click Continue.
   
   d. On the Add Service Wizard Select the set of dependencies for your new service page, select the dependencies for the service. Click Continue.

   Note: If the dependencies are automatically selected for this service, you might not see this page.

   e. On the Add Service Wizard Customize Role Assignments page, select a node for the service.

   • Choose any single node.

   • Kerberos: Select a node that has a Kerberos keytab for the hdfs user on that node.

   Click OK, and then click Continue.

   f. Enter your HDFS user name.

   • The default user name is hdfs.

   • Kerberos: A valid Kerberos ticket for your hdfs user must be available on the node that was selected for the SASQKB service.

   **TIP** Run kinit as the hdfs user to obtain a ticket, if necessary.

   g. Click Continue, and then click Finish.

   h. Verify that the QKB index file exists in the Hadoop file system: /sas/qkb/default.idx.

   i. If the SASQKB service that you just deployed is started, navigate to Cloudera Manager Home and stop the service.

5. Continue to “Return to Your Scenario for the Next Step” on page 119.

**Deploy the Stack Using Ambari**

1. Make sure that you have created the stack and copied it to the Hadoop cluster manager node, as described in “Create a Parcel or Stack of the QKB” on page 112.

2. Log on to Ambari.

3. Deploy the QKB to the Hadoop cluster:

   a. Click Actions and select + Add Service.

      The Add Service Wizard page and the Choose Services panel appear.
b. In the **Choose Services** panel, select the **SASQKB** service. Click **Next**.

The **Assign Slaves and Clients** panel appears.

c. In the **Assign Slaves and Clients** panel, select the nodes where you want the stack to be deployed.

**Note:**
- You must select the NameNode. Select any other nodes that might run applications submitted from SAS Data Loader for Hadoop or other SAS software.
- Depending on your version of Ambari, all nodes might be selected.

The **Customize Services** panel appears, and the SASQKB service stack is listed.

d. Do not change any settings on the **Customize Services** panel. Click **Next**.

**Note:** If your cluster is secured with Kerberos, the **Configure Identities** panel appears. Enter your Kerberos credentials in the **admin_principal** and **admin_password** text boxes.

If your cluster is secured with Kerberos, the **Configure Identities** panel appears. Enter your Kerberos credentials in the **admin_principal** and **admin_password** text boxes. Click **Next**.

The **Review** panel appears.

e. Review the information about the panel. If everything is correct, click **Deploy**.

The **Install, Start, and Test** panel appears. When the stack is installed on all nodes, click **Next**.

The **Summary** panel appears.

f. Click **Complete**. The stacks are now installed on all nodes of the cluster.

The SASQKB service is displayed on the Ambari dashboard.

g. Verify that the QKB index was created on the nodes: `/sas/qkb/default.idx`.

4. Continue to “Return to Your Scenario for the Next Step” on page 119.

---

**Using the QKB_PUSH.SH Script**

**Prerequisites**
- You can use the script to deploy the QKB on the DataNodes for all supported Hadoop distributions.
- Make sure that you know the location of the QKB, as described in “Prepare the SAS Quality Knowledge Base for Deployment” on page 29.
- If you want to replace a QKB on the Hadoop cluster, you must remove the old one first. For more information, see “Remove SAS Software from the Hadoop Cluster” on page 119.
- If you have already deployed the QKB by using Cloudera Manager or Ambari, do not perform the instructions in this topic.

The person who runs the qkb_push.sh script must have the following permissions:
- Passwordless SSH to the DataNodes from the NameNode.
Permission to use sudo to execute commands as the HDFS user or MapR user.

In a UNIX environment, terminals are referred to as TTYs. Hadoop administration frameworks, such as Cloudera Manager, might run scripts in a context that does not have a terminal. This might cause errors to occur in the qkb_push.sh script because it relies on the sudo command, which typically requires TTY-only access. A workaround is to temporarily disable the requiretty option in the user configuration file for the person who uses sudo to run the qkb_push.sh script.

- Write permission to the /opt folder on all nodes in the cluster. Alternatively, the root user must create a symlink from /opt/qkb to an area on the node to which you have Write permission.
- For an HDFS cluster, Write permission to the root of HDFS. For a MapR cluster, Write permission to the root of MapR-FS.
- For an HDFS cluster secured with Kerberos, the person who runs the qkb_push.sh script must use kinit to acquire a Kerberos ticket for himself or herself.
- Each Hadoop node might require up to 8 GB of disk space for the QKB.

**Copy the QKB to the Hadoop Cluster**

Copy the QKB installation folder to a staging area on the Hadoop NameNode. An example staging area might be: /tmp/qkbstage. You can use a file transfer command like FTP or SCP, or you can mount the file system where the QKB is located on the Hadoop NameNode.

The following example shows how you might copy a QKB that exists on a Linux system to the Hadoop NameNode. The example uses secure copy with the -r argument to recursively copy the specified directory and its subdirectories.

- Assume that hmaster456 is the host name of the Hadoop NameNode.
- The target location on the NameNode is /tmp/qkbstage

To copy the QKB from the client desktop, issue a command such as the following:

```
scp -r /opt/sas/qkb/ci/28 hmaster456:/tmp/qkbstage/
```

Your next task is to deploy the QKB to other nodes on the cluster.

**Run the Script**

If the QKB installation folder has been copied to the Hadoop NameNode, your next step is to deploy the QKB to all nodes on the Hadoop cluster. Deploy the QKB to all nodes that will run SAS Data Loader for Hadoop directives.

Installing the QKB on the Hadoop cluster nodes performs the following tasks:

- copies the specified QKB directory to a fixed location (/opt/qkb/default) on each of the Hadoop nodes.
- generates an index file from the contents of the QKB and pushes this index file to HDFS or MapR-FS. This index file, named default.idx, is created in the /sas/qkb directory in HDFS or MapR-FS. The default.idx file provides a list of QKB definition and token names to SAS Data Loader for Hadoop.

The qkb_push.sh script automatically discovers all nodes of the cluster by default and deploys the QKB to those nodes. Perform the following steps to run qkb_push.sh with the default options:

1. Verify that you have the required access privileges as described in “Prerequisites” on page 114.
2. Locate the qkb_push.sh script, which can be in one of the following directories:
   - `<EPInstallDir>/SASEPHome/bin`
   - `/opt/sas/ep/home/bin`
   - `/opt/sasep/home/bin`

   **Note:**
   - The qkb_push.sh script is deployed with the SAS Embedded Process software, and the directory where the script is located differs by the SAS Embedded Process release. For more information, see details about the different SAS Embedded Process releases on page 72 in the Hadoop administrator checklist.
   - The `ep-config.xml` file contains the path to the SAS Embedded Process home directory. The `ep-config.xml` file is located in the HDFS or MapR-FS directory: `/sas/ep/config/ep-config.xml`.

3. Run the script with a command such as the following:
   ```
   .qkb_push.sh /tmp/qkbstage/
   ```
   Alternatively, you can use the `-h` or `-f` arguments to specify deploying the files to a specific node or group of nodes.

   By default, qkb_push.sh does not list the names of the host nodes to which it deploys the files. To create such a list, include the `-v` argument in the command. If a name other than the default was configured for the HDFS or MapR user name, include the `-s` argument in the command.

   For information about supported arguments, see “Reference: QKB_PUSH.SH Syntax” on page 117.

   The qkb_push.sh script creates directories and files on each node on which it is executed. Verify that QKB directories and files have been copied to the nodes. The following directories are examples:
   ```
   /opt/qkb/default/chopinfo
   /opt/qkb/default/grammar
   /opt/qkb/default/locale
   /opt/qkb/default/phonetx
   /opt/qkb/default/regexlib
   /opt/qkb/default/scheme
   /opt/qkb/default/vocab
   ```
   Verify that the QKB index file (default.idx) was created on the Hadoop cluster by issuing the command:
   ```
   hadoop fs -ls /sas/qkb
   ```
   Continue to “Return to Your Scenario for the Next Step” on page 119.

**Reference: QKB_PUSH.SH Overview**

The qkb_push.sh script enables you to perform the following actions.

- Install or remove SAS QKB files on a single node or a group of nodes.
- Generate a SAS QKB index file and write the file to an HDFS or MapR-FS location.
- Write the installation or removal output to a log file.

The qkb_push.sh file is saved to the `<EPInstallDir>/SASEPHome/bin` directory. You must execute qkb_push.sh from this directory.
You can suppress index creation or perform only index creation by using the -x or -i arguments.

Note: Only one QKB and one index file are supported in the Hadoop framework at a time. Subsequent QKB and index pushes replace prior ones, unless you are pushing a QKB that is an earlier version than the one installed or has a different name. In these cases, you must remove the old QKB from the cluster before deploying the new one.

The QKB source directory is copied to the fixed location /opt/qkb/default on each node. The QKB index file is created in the /sas/qkb directory in HDFS or MapR-FS. If a QKB or QKB index file already exists in the target location, the new QKB or QKB index file overwrites it.

Reference: QKB_PUSH.SH Syntax
```
qkb_push.sh <arguments> qkb_path
```

Arguments
- `?` prints usage information.
- `-l logfile` directs status information to the specified log file instead of to standard output.
- `-f hostfile` specifies the full path of a file that contains the list of hosts where the QKB is installed or removed.
- `-h hostname` specifies the target host or host list where the QKB is installed or removed.

Default
The qkb_push.sh script discovers the cluster topology and uses the retrieved list of data nodes.

Interaction
Use the `-f` argument in conjunction with the `-r` argument to remove the QKB from specific nodes.

Note
The `-f` and `-h` arguments are mutually exclusive.

See
- `-r`

Example
- `-f /etc/hadoop/conf/slaves`

-h hostname < hostname>
specifies the target host or host list where the QKB is installed or removed.

Default
The qkb_push.sh script discovers the cluster topology and uses the retrieved list of data nodes.
<table>
<thead>
<tr>
<th>Requirement</th>
<th>If you specify more than one host, the host names must be separated by spaces.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interaction</td>
<td>Use the -h argument in conjunction with the -r argument to remove the QKB from specific nodes.</td>
</tr>
<tr>
<td>Note</td>
<td>The -f and -h arguments are mutually exclusive.</td>
</tr>
<tr>
<td>Tip</td>
<td>Use the -host argument when new nodes are added to the cluster</td>
</tr>
<tr>
<td>See</td>
<td>-r</td>
</tr>
</tbody>
</table>
| Example     | `-h server1 server2 server3`  
|             | `-h bluesvr` |

`-v`

specifies verbose output, which lists the names of the nodes on which the script ran.

`-s user-id`

specifies the user ID that has Write access to the HDFS or MapR-FS root directory when the default user name is not used.

| Defaults | hdfs for all Hadoop distributions except MapR  
|          | mapr for MapR |

`-i`

creates and pushes the QKB index only.

`-I`

generates the QKB index file to a specified path on your local filesystem.

`-x`

suppresses QKB index creation.

`-r`

removes the QKB from the Hadoop nodes and it removes the QKB index file from HDFS or MapR-FS.

| Default | The -r argument discovers the cluster topology and uses the retrieved list of data nodes. |
| Interaction | You can specify the hosts from which you want to remove the QKB by using the -f or -h arguments. The -f and -h arguments are mutually exclusive. |
| Note | The QKB index file is not removed from HDFS or MapR-FS when the -h or -f argument is specified in conjunction with -r. |
| See | `-f hostfile`  
|     | `-h hostname hostname` |
| Example | `-r -h server1 server2 server3`  
|         | `-r -f /etc/hadoop/conf/slaves`  
|         | `-r -l logfile` |
Remove SAS Software from the Hadoop Cluster

Why You Might Remove SAS Software from the Hadoop Cluster

Key features of SAS Data Loader for Hadoop depend on the following SAS software to be installed on the Hadoop cluster:

- SAS Embedded Process
- SAS Data Loader for Hadoop Spark Engine
- SAS Quality Knowledge Base (QKB)

You should remove this SAS software from the Hadoop cluster for the following reasons:

- You want to update the software listed above to a later version.
- You no longer use a Hadoop cluster with SAS Data Loader for Hadoop (for example, if you switch to a different Hadoop distribution).
- You want to replace software, such as replacing an old QKB with a new one.

Guidance for Removing SAS Software

The steps to remove the SAS software depend on how the software was installed. Refer to the following table to understand the different methods.

<table>
<thead>
<tr>
<th>How Software Was Deployed</th>
<th>Documentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Used shell scripts to deploy</td>
<td>“Remove Software Using Shell Scripts” on page 120</td>
</tr>
</tbody>
</table>
How Software Was Deployed | Documentation
---|---
Deployed parcels using Cloudera Manager | “Remove Parcels Using Cloudera Manager” on page 122

Deployed stacks using Ambari | “Remove Stacks Using Ambari” on page 124

**Important:** The instructions to remove the SAS Embedded Process differ by release.

- For more information, see the details about the different SAS Embedded Process releases on page 72 in the Hadoop administrator checklist.
- If the instructions in this guide were used to deploy the SAS Embedded Process, then use the instructions in this topic to remove the SAS Embedded Process.
- If the instructions in the SAS Embedded Process: Deployment Guide were used to deploy the SAS Embedded Process, then use the instructions in that guide to remove the SAS Embedded Process.

**Remove Software Using Shell Scripts**

**CAUTION:**

Remove in the order shown if you plan to remove all software shown below. If a QKB is installed, remove it first, and then remove the SAS Data Loader for Hadoop Spark Engine before removing the SAS Embedded Process.
Run the `qkb_push.sh` script with the `-r` option. Here is an example:

```bash
cd EPInstallDir/saseexe/SASEPHome/bin/.
./qkb_push.sh -r
```

The `qkb_push.sh` script is deployed with the SAS Embedded Process software, and the directory path shown differs for each release. For more information, see details about the different SAS Embedded Process releases on page 72 in the Hadoop administrator checklist.

Locate the `qkb_push.sh` script in one of the following directories:

- `<EPInstallDir>/SASEPHome/bin`
- `/opt/sas/ep/home/bin`
- `/opt/sasep/home/bin`

The `ep-config.xml` file contains the path to the SAS Embedded Process home directory. The `ep-config.xml` file is located in the HDFS or MapR-FS directory: `/sas/ep/config/ep-config.xml`.

To remove the QKB, make sure that you have the required credentials on page 114.

The `-r` argument automatically discovers all nodes of the cluster by default and removes the QKB files from those nodes. Use the `-h` or `-f` arguments to specify removing the files from a specific node or group of nodes.

The QKB index file is not removed from HDFS when the `-h` or `-f` argument is specified with `-r`.

By default, the `-r` argument does not list the names of the host nodes from which it removes the files. To create such a list, include the `-v` argument in the command.

For information about supported arguments, see “Reference: QKB_PUSH.SH Overview” on page 116.

---

Run the `sasdmp-admin.sh` script with the `-remove` option.

```bash
cd DMPInstallDir/saseexe/SASDMPHome/bin/.
sasdmp-admin.sh -remove
```

To remove the SAS Data Loader for Hadoop Spark Engine, make sure that you have the required credentials on page 101.

For information about supported arguments, see “Reference: SASDMP_ADMIN.SH Syntax” on page 104.
Run the sasep-admin.sh script with the -remove option.

cd EPInstallDir/sasexe/SASEPHome/bin/
./sasep-admin.sh -remove

The directory path shown pertains to a specific release of the SAS Embedded Process. Depending on the release of the SAS Embedded Process that was deployed, the location of the sasep-admin.sh script differs. For more information, see details about the different SAS Embedded Process releases on page 72 in the Hadoop administrator checklist.

Locate the sasep-admin.sh script in one of the following directories:

• <EPInstallDir>/SASEPHome/bin
• /opt/sas/ep/home/bin
• /opt/sasep/home/bin

The ep-config.xml file contains the path to the SAS Embedded Process home directory. The ep-config.xml file is located in the HDFS or MapR-FS directory: /sas/ep/config/ep-config.xml.

To remove the SAS Embedded Process, make sure that you have the required credentials on page 82.

For information about supported arguments, see “Reference: SASEP-ADMIN.SH Syntax” on page 87.

---

### Remove Parcels Using Cloudera Manager

#### Overview

Consider the following tips when working with parcels:

- If you install a new version or re-install a previous version of a parcel, deactivate or remove the existing parcel after installing the new one.
- You can have more than one parcel for a particular product on the Hadoop cluster, but only one can be active.
- Deactivate parcels that you do not want to use. You can also remove them from the Hadoop cluster after deactivation.

#### Deactivating Parcels

*Note:* Deactivation and removal of the parcels for the SAS Data Loader for Hadoop Spark Engine and the SAS Embedded Process follow the same steps. The parcel names are SASDMSPARK and SASEP, respectively. These names are represented in the following procedures by parcel_name. The configuration filenames for the SAS Data Loader for Hadoop Spark Engine and the SAS Embedded Process are dmp-config.xml and ep-config.xml, respectively. These filenames are represented in the following procedures by file_name.

To deactivate a parcel using Cloudera Manager, follow these steps:
1. Log on to Cloudera Manager.

2. If running, stop any of the parcel_name services:
   a. On the Home page, click the down arrow next to parcel_name service.
   b. Under parcel_name Actions, select Stop, and then click Stop.

3. Delete the parcel_name service from Cloudera Manager:
   Note: If you are deleting more than one service, delete all services before proceeding to the step of deactivation.
   a. On the Home page, click the down arrow next to parcel_name service.
   b. Click Delete. The parcel_name service no longer appears on the Home ⇒ Status tab.

4. Deactivate the parcel_name parcel:
   a. Navigate to the Hosts ⇒ Parcels tab.
   b. For parcel_name, select Actions ⇒ Deactivate. You are prompted either to restart the cluster or close the window.
   c. When prompted, click Close.
   CAUTION:
   Do not restart the cluster.
   d. Click OK to continue the deactivation.

Removing Parcels
After deactivating the parcel, follow these steps to remove it:

1. Remove the parcel_name parcel:
   a. For parcel_name, select Activate ⇒ Remove from Hosts.
   b. Click OK to confirm.

2. For parcel_name, select Distribute ⇒ Delete.

3. Click OK to confirm.
   This step deletes the parcel files from the /opt/cloudera/parcel directory.

4. Manually remove the file_name file:
   a. Log on to HDFS.

      sudo su - root
      su - hdfs | hdfs-userid

      Note: If your cluster is secured with Kerberos, the HDFS user must have a valid Kerberos ticket to access HDFS. This can be done with kinit.

   b. Navigate to the appropriate directory on HDFS. The directory for SASEP and SASDMSpark is /sas/ep/config/.
   c. Delete the file_name file.
**Remove Stacks Using Ambari**

**Overview**
Consider the following tips when working with stacks:

- If you install a new version or re-install a previous version of a stack, deactivate or remove the existing stack after installing the new one.
- You can have more than one stack for a particular product on the Hadoop cluster, but only one can be active.
- Deactivate stacks that you do not want to use. You can also remove them from the Hadoop cluster after deactivation.

**Deactivating Stacks**
*Note:* Deactivation and removal of the stacks for the SAS Data Loader for Hadoop Spark Engine and the SAS Embedded Process follow the same steps. The stack names are SASDMSPARK and SASEP, respectively. These names are represented in the following procedures by `stack_service`. The configuration filenames for the SAS Data Loader for Hadoop Spark Engine and the SAS Embedded Process are `dmp-config.xml` and `ep-config.xml`, respectively. These filenames are represented in the following procedures by `file_name`.

You deactivate a stack by activating another stack.

To deactivate a stack using Ambari, follow these steps:

1. Log on to the Ambari manager. All deployed versions of the `stack_service` stack appear in the left pane of the Home page under the `allversions` text box.
2. Select the `stack_service` stack that you want to activate.
3. Enter the version number of the stack that you want to activate in the `activated_version` text box on the Configs tab.
4. Click Save.
5. You add a note describing your action, and then click OK.
6. If you are deactivating more than one stack, finish all deactivation tasks before restarting services.
7. Click Restart to restart the `stack_service` after you have deactivated all the stacks.
8. Click Restart All Affected. The affected services are restarted.
9. The new stack is activated, leaving the previous stack deactivated.
10. If you have deactivated additional stacks, select them and restart all affected services. The new stacks are activated, leaving the previous stacks deactivated.

**Removing Stacks**
*Note:* Root or passwordless sudo access is required to remove the stack.

After deactivating the stack, follow these steps to remove it:

1. Navigate to the appropriate `Admin/bin/stack` directory, where `stack` represents either `sasep` or `dmspark`. These directories are on the Linux machine where SAS software on the cluster is downloaded and unzipped.

   A `delete_stack.sh` file is in each `stack` directory.
2. Copy the `delete_stack.sh` file to a temporary directory where the cluster manager server is located. Here is an example using secure copy.

```bash
scp delete_stack.sh user@cluster-manager-host:/mytemp
```

3. Use this command to run the delete script.

```bash
./delete_stack.sh Ambari-Admin-User-Name
```

4. Enter the Ambari administrator password at the prompt.

A message appears that offers options for removal.

5. Enter one of the options:

- Enter 1 to remove only the `file_name` file.
- Enter 2 to remove a specific version of `stack_service`.
- Enter 3 to remove all versions of `stack_service`.

You are prompted to restart the Ambari server to complete the removal of the `stack_service`.

6. Enter `y` to restart the Ambari server. The `stack_service` no longer appears.

---

**Deploy the SAS Data Loader for Hadoop Spark Engine Using SAS Deployment Manager**

**Before You Begin**

The **Deploy SAS Data Loader for Hadoop Spark Engine** task in the SAS Deployment Manager can generate a Cloudera parcel or a Hortonworks stack for the engine. This task copies the parcel or stack to the cluster manager node on your Hadoop cluster. The cluster manager application (Cloudera Manager or Ambari) can then be used to deploy the engine to the DataNodes on the Hadoop cluster. This combination of tools and credentials requires close collaboration between the SAS administrator and the Hadoop administrator.

**Note:**

To use the **Deploy SAS Data Loader for Hadoop Spark Engine** task in the SAS Deployment Manager:

- Your Hadoop cluster must be based on UNIX.
- Your SAS server tier must be based on UNIX. The Spark engine task in the SAS Deployment Manager is available only on UNIX.
- Your Hadoop distribution must be a supported version of Cloudera CDH or Hortonworks.
- You must obtain the following information:
  - the host name and port number for your Hadoop cluster manager host
  - a UNIX user account with SSH for the cluster manager host
  - credentials for the cluster manager application (Cloudera Manager or Ambari): this is a UNIX account name that has administrator privileges and a password for the cluster manager software
the path to the `standalone_installs` directory; this is a subdirectory of the SAS Software Depot that contains the installation files for the SAS Data Loader for Hadoop Spark Engine

instructions for logging on to the cluster manager software

To use Cloudera Manager or Ambari to deploy the SAS Data Loader for Hadoop Spark Engine:

• You must have Write permission to the root of HDFS.
• If your Hadoop cluster is secured with Kerberos, use `kinit` to acquire a Kerberos ticket for the HDFS user who runs the cluster manager.

If you want to replace the SAS Data Loader for Hadoop Spark Engine software on the Hadoop cluster, you must remove the old software first. For more information, see “Remove SAS Software from the Hadoop Cluster” on page 119.

**Generate a Parcel or Stack Using SAS Deployment Manager**

Perform the following steps to run the **Deploy SAS Data Loader for Hadoop Spark Engine** task in the SAS Deployment Manager:

1. Open a command prompt on a machine in the SAS server tier.
2. Navigate to the SAS Deployment Manager directory. An example path might be: `<SASHome>/SASDeploymentManager/9.4`
3. On UNIX, enter `.\sasadm.sh`.
4. Select a language and click **OK**. The first page of the SAS Deployment Manager is displayed.
5. Select the **Deploy SAS Data Loader for Hadoop Spark Engine** task and click **Next**.
6. Select your Hadoop distribution (Cloudera or Hortonworks) and click **Next**.
7. Enter the host name and port number for your Hadoop cluster manager host machine. Click **Next**.
8. Enter credentials for the Hadoop cluster manager application. This is a UNIX account name that has administrator privileges and a password for your cluster manager software (Cloudera Manager or Ambari). Click **Next**.
9. Enter credentials for the Hadoop cluster manager host. This is a UNIX user account with SSH for the Hadoop cluster manager host machine. Click **Next**.
10. Specify the path to the `standalone_installs` directory. This is a subdirectory of the SAS Software Depot that contains the installation files for the SAS Data Loader for Hadoop Spark Engine.
11. Accept the default temporary installation directory or specify a new one. Click **Next**.
12. Allow the SAS Deployment Wizard to verify that the installation can proceed. Click **Next**.
13. Review the status of the SAS Data Loader for Hadoop Spark Engine deployment. If all is well, click **Start**.
14. The task generates a Cloudera parcel or a Hortonworks stack for the SAS Data Loader for Hadoop Spark Engine. The task copies the parcel or stack to the cluster manager node on your Hadoop cluster.
Your next task is to deploy the engine to the DataNodes on the Hadoop cluster. Use one of the following methods: “Deploy the Parcel Using Cloudera Manager” on page 127 or “Deploy the Stack by Using Ambari” on page 128.

**Deploy the Parcel Using Cloudera Manager**

Before you begin, make sure that you have created the parcel of the SAS Data Loader for Hadoop Spark Engine. For more information, see “Generate a Parcel or Stack Using SAS Deployment Manager” on page 126.

1. Log on to Cloudera Manager.
2. Perform the following steps:
   a. From the Menu bar, select **Hosts** → **Parcels**.  
      **Note:** If the SASDMSPARK parcel is missing, run **Check for new parcel**.
   b. On the row for the SASDMSPARK parcel, click **Distribute** to copy the parcel to all nodes.
   c. Click **Activate**.
      **CAUTION:** Do not restart the cluster.
      After you complete the activation, you might be prompted to either restart the cluster or to close the window.
   d. If prompted, click **Close**.
3. Add the SASDMSPARK service to create the SASDMSPARK configuration file in HDFS.
   a. Navigate to Cloudera Manager Home.
   b. In Cloudera Manager, select the drop-down arrow next to the name of the cluster, and then select **Add a Service**. The Add Service Wizard appears.
   c. Select the SASDMSPARK service, and then click **Continue**.
   d. On the **Add Service Wizard** → **Select the set of dependencies for your new service** page, select the dependencies for the service. Click **Continue**.  
      **Note:** If the dependencies are automatically selected for this service, you might not see this page.
   e. On the **Add Service Wizard** → **Customize Role Assignments** page, select a node for the service.
      • Choose any single node.
      • **Kerberos:** Select a node that has a Kerberos keytab for the hdfs user on that node.
      Click **OK**, and then click **Continue**.
   f. Enter your HDFS user name.
      • The default user name is hdfs.
      • **Kerberos:** A valid Kerberos ticket for your hdfs user must be available on the node that was selected for the SASDMSPARK service.
      **TIP** Run kinit as the hdfs user to obtain a ticket, if necessary.
   g. Click **Continue**, and then click **Finish**.
h. Verify that the configuration file /sas/ep/config/dmp-config.xml is added to the HDFS file system.

i. If the SASDMSPARK service that you just deployed is started, navigate to Cloudera Manager Home and stop the service.

**Deploy the Stack by Using Ambari**

Before you begin, make sure that you have created the stack of the SAS Data Loader for Hadoop Spark Engine. For more information, see “Generate a Parcel or Stack Using SAS Deployment Manager” on page 126.

1. Log on to Ambari.

2. Perform the following steps:
   a. Click **Actions** and select + **Add Service**.
      The **Add Service Wizard** page and the **Choose Services** panel appear.
   
   b. In the **Choose Services** panel, select the **SASDMSPARK** service. Click **Next**.
      The **Assign Slaves and Clients** panel appears.
   
   c. In the **Assign Slaves and Clients** panel, select the nodes where you want the stack to be deployed.
      
      **Note:**
      - You must select the NameNode. Select any other nodes that might run applications submitted from SAS Data Loader for Hadoop or other SAS software.
      - Depending on your version of Ambari, all nodes might be selected.

      The **Customize Services** panel appears, and the SASDMSPARK service stack is listed.
   
   d. Do not change any settings on the **Customize Services** panel. Click **Next**.
      
      **Note:** If your cluster is secured with Kerberos, the **Configure Identities** panel appears. Enter your Kerberos credentials in the **admin_principal** and **admin_password** text boxes.

      If your cluster is secured with Kerberos, the **Configure Identities** panel appears. Enter your Kerberos credentials in the **admin_principal** and **admin_password** text boxes. Click **Next**.

      The **Review** panel appears.
   
   e. Review the information about the panel. If everything is correct, click **Deploy**.
      
      The **Install, Start, and Test** panel appears. When the stack is installed on all nodes, click **Next**.

      The **Summary** panel appears.
   
   f. Click **Complete**. The stacks are now installed on all nodes of the cluster.

      The SASDMSPARK service is displayed on the Ambari dashboard.
   
   g. Verify that the configuration file /sas/ep/config/dmp-config.xml is added to the HDFS file system.
Recommended Reading

- *SAS Data Loader for Hadoop: User’s Guide*

For a complete list of SAS publications, go to sas.com/store/books. If you have questions about which titles you need, please contact a SAS Representative:

SAS Books
SAS Campus Drive
Cary, NC 27513-2414
Phone: 1-800-727-0025
Fax: 1-919-677-4444
Email: sasbook@sas.com
Web address: sas.com/store/books
Gain Greater Insight into Your SAS® Software with SAS Books.

Discover all that you need on your journey to knowledge and empowerment.

support.sas.com/bookstore for additional books and resources.