SAS® Contextual Analysis In-Database Scoring 14.3 for Hadoop: Administrator's Guide

SAS® Documentation
August 28, 2017

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2017. *SAS® Contextual Analysis In-Database Scoring 14.3 for Hadoop: Administrator's Guide*. Cary, NC: SAS Institute Inc.

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

September 2017

14.3-P1:ctxtindbag

Contents

SAS Contextual Analysis In-Database Scoring 14.3 for Hadoop: Administrator's Guide

Audience

This book is for users of SAS Contextual Analysis who want to use their scoring models in Hadoop. The book provides instructions for deploying the SAS In-Database Scoring for Hadoop and also deploying your models the Hadoop environment. It is assumed that you know how to use SAS Contextual Analysis.

Requirements

You must have the following products licensed:

- SAS Contextual Analysis
- SAS In-Database Code Accelerator for Hadoop

Chapter 1

Introduction

Overview of SAS Contextual Analysis In-Database Scoring for Hadoop

Enterprises store large amounts of unstructured text documents, along with other data, in Hadoop and are seeking tools that enable the analysis of all data within the database in order to avoid excessive consumption of network resources. SAS Contextual Analysis In-Database Scoring for Hadoop enables IT professionals to deploy text analytics models (also called binary models) inside their Hadoop infrastructure, which avoids any data movement during scoring and takes advantage of existing data warehouse investments.

About SAS Contextual Analysis

SAS Contextual Analysis is a web-based application that uses natural language processing and machine learning to derive insights from textual data through topic identification, categorization, entity and fact extraction, and sentiment analysis—all from a single interface. Using this application, you can build models (based on training documents) and create taxonomies and rule sets to analyze documents. You can then customize your models for your business domain in order to realize the value of your text-based data.

At the end of the modeling process, SAS Contextual Analysis generates DS2 code and binary models for scoring text data. The score code can be retrieved from the View drop-down menu, which is first seen on the Properties page of any SAS Contextual Analysis project.

SAS Contextual Analysis generates three types of DS2 score code and models corresponding to categorization, concept extraction, and sentiment analysis. The DS2 code can be run within a SAS environment such as SAS Studio or modified to run within your Hadoop cluster using SAS In-Database Code Accelerator for Hadoop. The binary models represent the rule sets for categorization (file extension: .mco), concepts (file extension: .li) and sentiment (file extension: .sam) taxonomies, which are highly optimized to apply all rules in parallel.

For information about using SAS In-Database Code Accelerator for Hadoop, see *SAS In-Database Products: User's Guide*.

What is SAS Embedded Process?

SAS Embedded Process is the core of SAS in-database products. It allows the parallel execution of SAS processes inside Hadoop or inside other databases. SAS Embedded Process technology is a portable, lightweight, execution container for SAS DS2 code. SAS Embedded Process is orchestrated by Hadoop MapReduce framework. Load balancing and resources allocation are managed by YARN.

SAS Embedded Process offers a flexible, efficient way to leverage increasing amounts of data by injecting the processing power of SAS where ever the data lives. SAS Embedded Process can tap into the massively parallel processing (MPP) architecture of Hadoop for scalable performance. Using SAS in-database technologies for Hadoop, you can run scoring models generated by SAS Contextual Analysis.

For more information about using SAS Embedded Process, see *SAS In-Database Products: Administrator's Guide*

Why Score Text Analytics Models In-Database?

The SAS In-Database Code Accelerator for Hadoop enables you to publish a DS2 multi-threaded program and its associated files to the database and execute that threaded program in parallel within SAS Embedded Process. Benefits of in-database processing include reduced data movement, faster run time, and the ability to leverage existing data warehousing investments without having to copy data to a secondary location for processing. Examples of threaded programs include large transpositions, computationally complex programs, scoring models, and BY-group processing.

Which Hadoop Platforms Are Supported?

The Cloudera and Hortonworks platforms are supported for this product.

Getting Started with the SAS Contextual Analysis In-Database Scoring for Hadoop

The steps for getting started with this product are as follows:

1. Deploy the In-Database Deployment Package for Hadoop. You do this by using the SAS Deployment Manager. For detailed steps, see Chapter 2, "Deploying the In-Database Deployment Package Using the SAS Deployment Manager," on page 5.
2. Deploy the SAS Contextual Analysis scoring models. For details, see Chapter 3, "Deploying SAS Text Analytics Scoring Models ," on page 27.

For information about using the models, see *SAS Contextual Analysis In-Database Scoring for Hadoop: User's Guide*.

Chapter 2

Deploying the In-Database Deployment Package Using the SAS Deployment Manager

When to Deploy the SAS In-Database Deployment Package Using the SAS Deployment Manager

You can use the SAS Deployment Manager to deploy the SAS In-Database Deployment Package if the following conditions are met:

- For Cloudera:
  - You are using Cloudera 5.8 or later. For the latest information, see the SAS Foundation system requirements documentation for your operating environment.
  - Cloudera Manager is installed.
  - Your other SAS software, such as Base SAS and SAS/ACCESS Interface to Hadoop, was installed on a UNIX server.
- For Hortonworks, IBM BigInsights, or Pivotal HD:
  - You are using Hortonworks 2.5, IBM BigInsights 4.2, or Pivotal HD 3.0 or later. For the latest information, see the SAS Foundation system requirements documentation for your operating environment.
  - You are using Ambari 2.4 or later.

- Your other SAS software, such as Base SAS and SAS/ACCESS Interface to Hadoop, was installed on a UNIX server.

Otherwise, you should deploy the SAS In-Database deployment package manually. For more information, see "Deploying the In-Database Deployment Package Manually" in *SAS In-Database Products: Administrator's Guide*.

CAUTION:

Once you have chosen a deployment method, you should continue to use that same deployment method when upgrading or redeploying SAS software on the cluster. Otherwise, your SAS software on the cluster can become unusable. For example, if you use the SAS Deployment Manager to deploy the SAS software on the cluster, you should continue to use the SAS Deployment Manager for upgrades or redeployments. You should not use the manual deployment method to upgrade or redeploy. If you do need to change deployment methods, you must first uninstall the SAS software on the cluster using the same method that you used to deploy it. You can then use the other deployment method to install it.

Prerequisites for Using the SAS Deployment Manager to Deploy the In-Database Deployment Package

The following prerequisites must be met before you can use the SAS Deployment Manager:

- You must have passwordless SSH access from the master node to the slave nodes.
- If your cluster is secured with Kerberos, in addition to having a valid ticket on the client, a Kerberos ticket must be valid on node that is running Hive. This is the node that you specify when using the SAS Deployment Manager.
- If you are using Cloudera, the SSH account must have Write permission to these directories:

  /opt/cloudera

  /opt/cloudera/csd

  /opt/cloudera/parcels

- You cannot customize the install location of the SAS Embedded Process on the cluster. By default, the SAS Deployment Manager deploys the SAS Embedded Process in the `/opt/cloudera/parcels` directory for Cloudera and the `/opt/sasep_stack` directory for Hortonworks, IBM BigInsights, and Pivotal HD.
- If you are using Cloudera, the Java JAR and GZIP commands must be available.
- If you are using Hortonworks, the requiretty option is enabled, and the SAS Embedded Process is installed using the SAS Deployment Manager, the Ambari server must be restarted after deployment. Otherwise, the SASEP service does not appear in the Ambari list of services. It is recommended that you disable the requiretty option until the deployment is complete.
- The following information is required:
  - host name and port of the cluster manager
  - credentials (account name and password) for the Hadoop cluster manager

- Hive service host name
- Oozie service host name (if required by your software)
- Impala service host name (if required by your software)
- credentials of the UNIX user account with SSH for the Hadoop cluster manager

Hadoop Installation and Configuration Steps Using the SAS Deployment Manager

To install and configure Hadoop using the SAS Deployment Manager, you must follow and complete these steps:

Note: If you are running both SAS 9.4 and SAS Viya, only the latest version of the SAS Embedded Process should be installed. For example, assume that you installed the SAS Embedded Process using the sepcorehadp-13.00000-1.sh file. Now you want to install SAS Viya and the in-database deployment package that shipped with SAS Viya, which is sepcorehadp-11.50000-1.sh. In this instance, you would not install the in-database deployment package that shipped with SAS Viya because it is an earlier version (that is, 11.50000 versus 13.0000).

Step	Description	Where to Go for Information
1	Review these topics:	• "Prerequisites for Installing the In-Database Deployment Package for Hadoop" in *SAS In-Database Products: Administrator's Guide* • "Backward Compatibility" in *SAS In-Database Products: Administrator's Guide* • "When to Deploy the SAS In-Database Deployment Package Using the SAS Deployment Manager" on page 5 • "Prerequisites for Using the SAS Deployment Manager to Deploy the In-Database Deployment Package" on page 6
2	If you have not already done so, configure SAS/ACCESS Interface to Hadoop. One of the key tasks in this step is to configure the Hadoop client files.	"Configuring SAS/ACCESS for Hadoop" in SAS Hadoop Configuration Guide for Base SAS and SAS/ACCESS
3	If you are upgrading from or re-installing a previous release, follow these instructions:	"Upgrading from or Re-installing a Previous Version If Using SAS Deployment Manager" in *SAS In-Database Products: Administrator's Guide*
4	Deploy the SAS Embedded Process parcel (Cloudera) or stack (Hortonworks, IBM BigInsights, or Pivotal HD) to the cluster.	For more information, see "Using the SAS Deployment Manager to Deploy the SAS Embedded Process Parcel or Stack to the Cluster" on page 8.

Step	Description	Where to Go for Information
5	Deploy the parcel (Cloudera) or stack (Hortonworks, IBM BigInsights, or Pivotal HD) to all the nodes on the cluster.	For more information see "Deploying the SAS Embedded Process Parcel on Cloudera" on page 20 or "Deploying the SAS Embedded Process Stack on Hortonworks, IBM BigInsights, or Pivotal HD" on page 22.
6	Review any additional configuration that might be needed depending on your Hadoop distribution.	For more information, see "Additional Configuration for the SAS Embedded Process" in *SAS In-Database Products: Administrator's Guide*.

Using the SAS Deployment Manager to Deploy the SAS Embedded Process Parcel or Stack to the Cluster

1. Start the SAS Deployment Manager by running sasdm.sh for UNIX. The SAS Deployment Manager script is located in the **`/SASHome/SASDeploymentManager/9.4/`** directory.

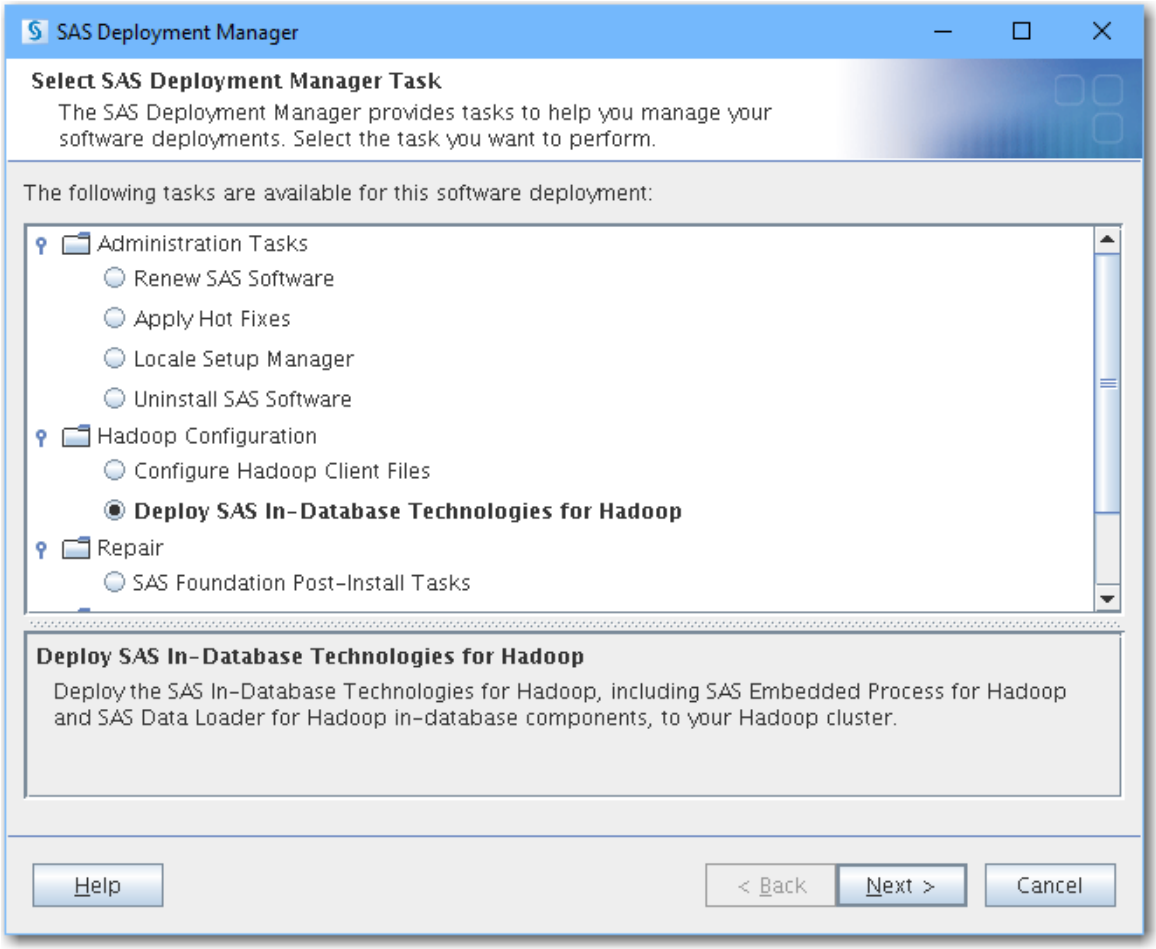
   ```
   ./sasdm.sh
   ```

   Note: For more information about the SAS Deployment Manager pages, click **Help** on each page.

   The **Choose Language** page opens.

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

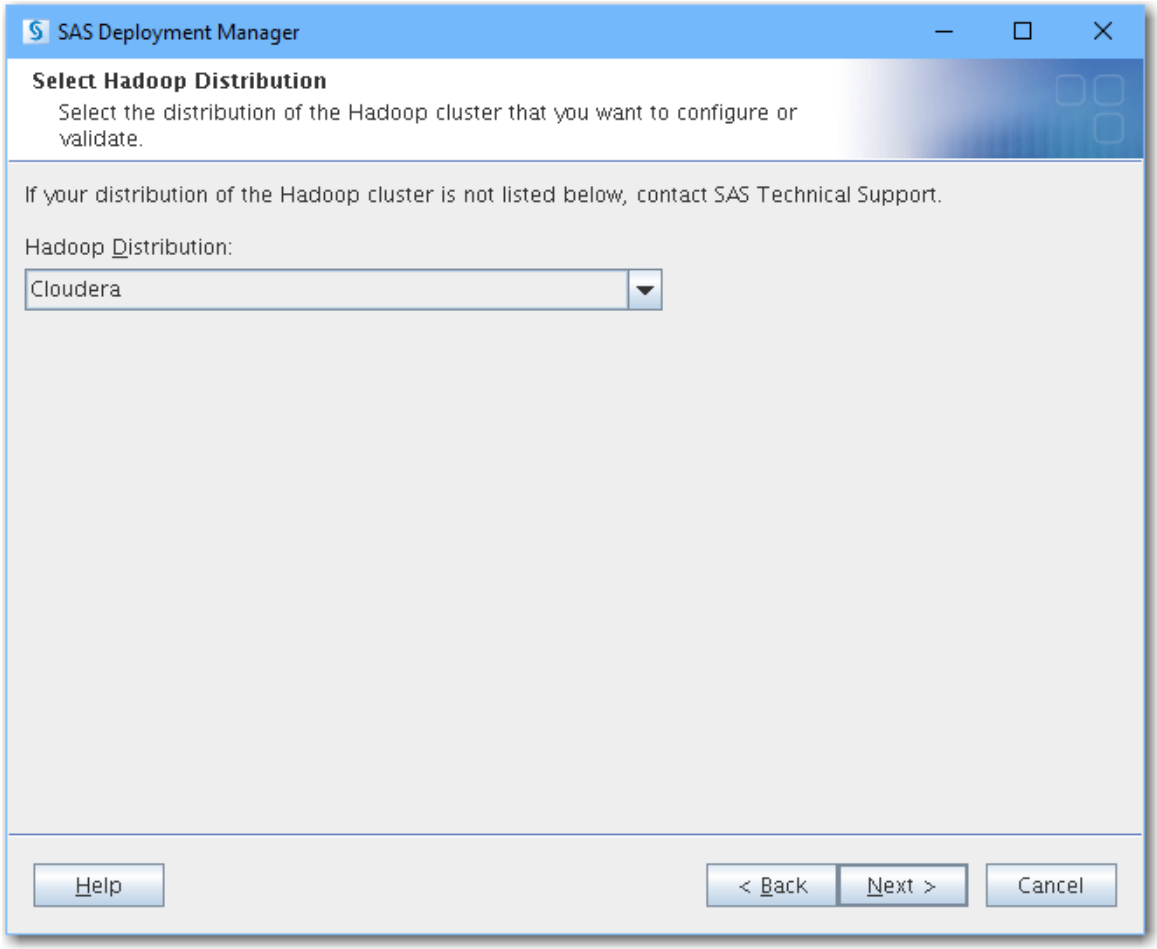
   Click **OK**. The **Select SAS Deployment Manager Task** page opens.



3. Under Hadoop Configuration, select **Deploy SAS In-Database Technologies for Hadoop**.

   Click **Next** to continue. The **Select Hadoop Distribution** page opens.

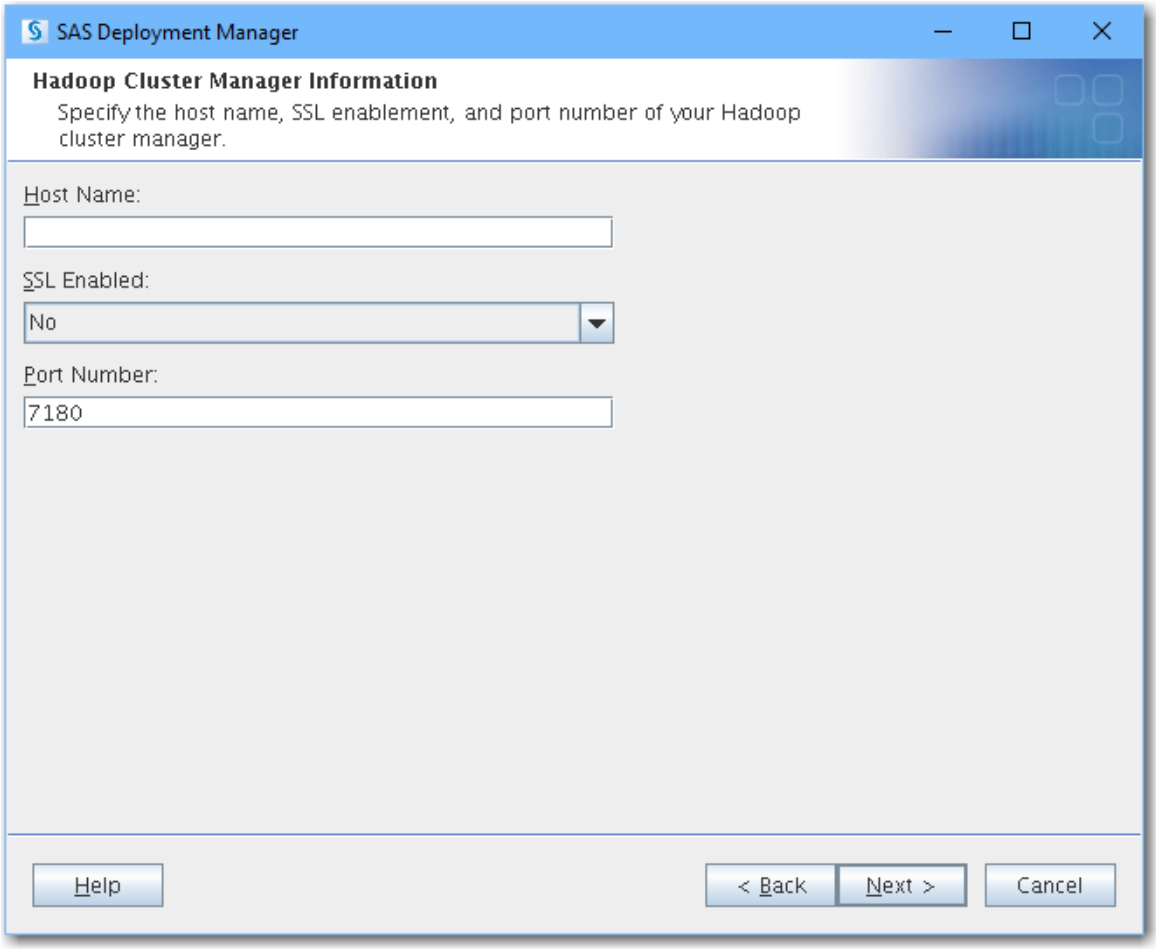
   Note: If you have licensed and downloaded SAS Contextual Analysis In-Database Scoring for Hadoop, the SAS Contextual Analysis In-Database Scoring for Hadoop component is silently deployed at the same time as the SAS Embedded Process for Hadoop.



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

   Note: If your distribution is not listed, exit the SAS Deployment Manager and contact SAS Technical Support.

   Click **Next**. The **Hadoop Cluster Manager Information** page opens.



5. 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 in Step 4 on page 10.

     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 it issues an error message if the URL is not accessible.

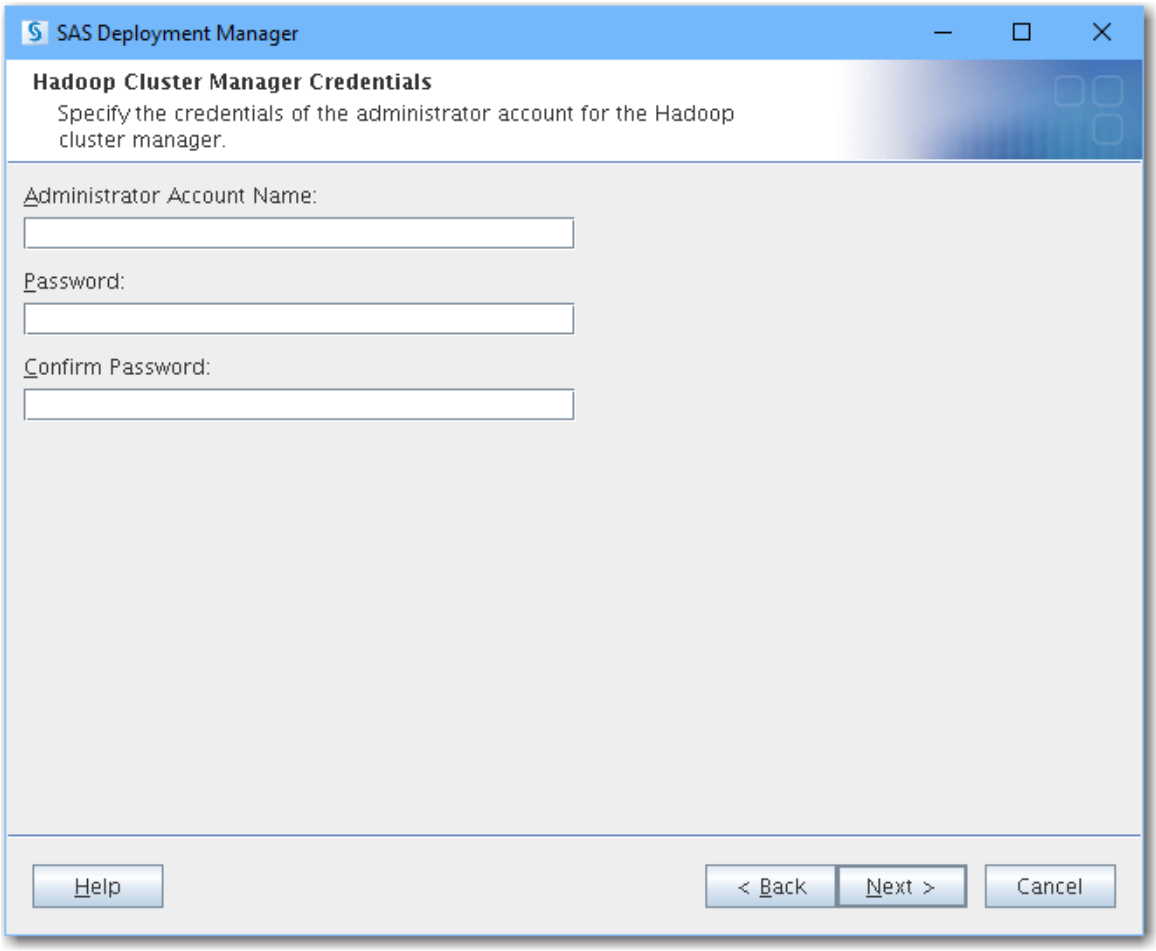
   - Indicate whether the cluster manager is enabled with Transport Layer Security (TLS).

     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 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 **Deploy SAS In-Database Technologies for Hadoop** task.

Click **Next**. The **Hadoop Cluster Manager Credentials** page opens.

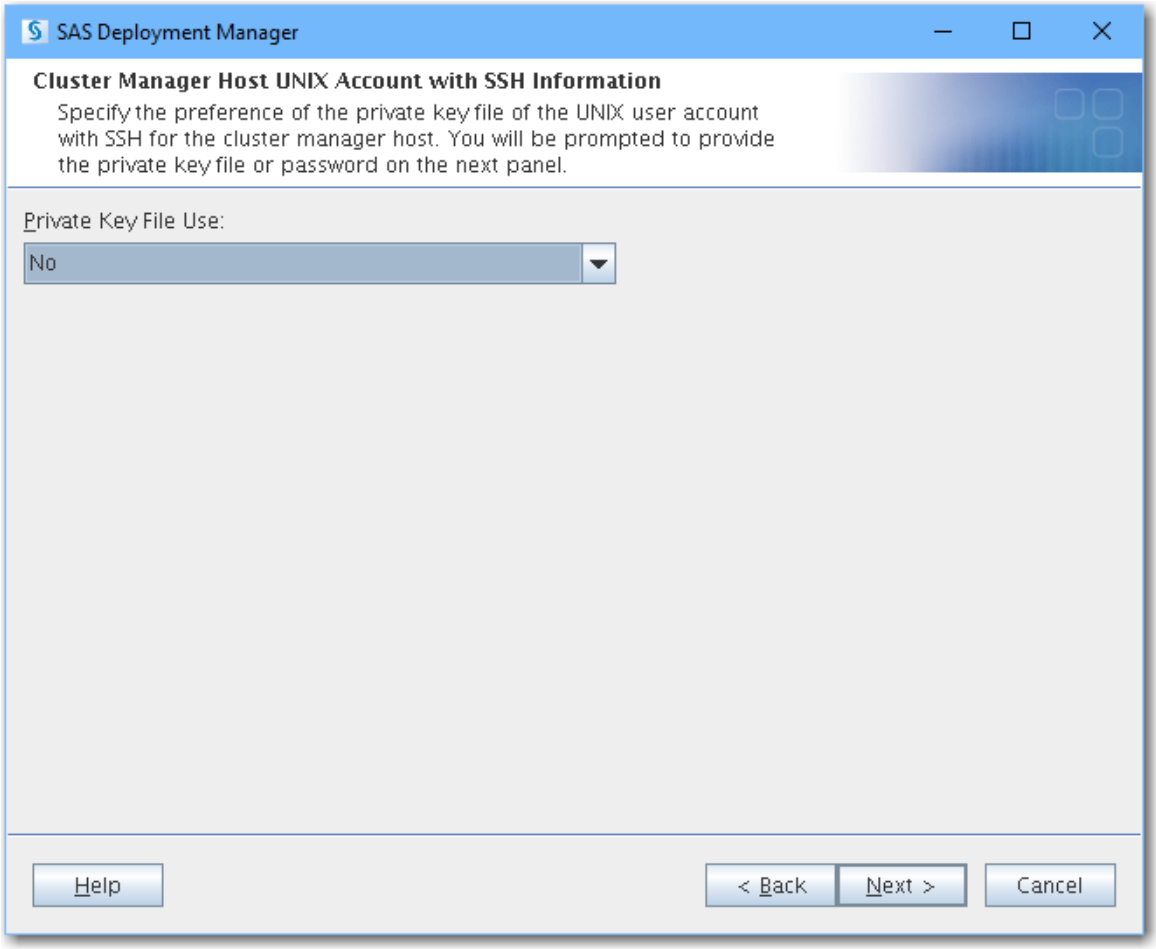


6. Enter the Cloudera Manager or Ambari administrator account name and password.

   Note: Using the credentials of the administrator account to query the Hadoop cluster and to find the Hive node eliminates guesswork and removes the chance of a configuration error. However, the account name does not have to be that of an administrator; it can be a read-only user.

   Click **Next**.

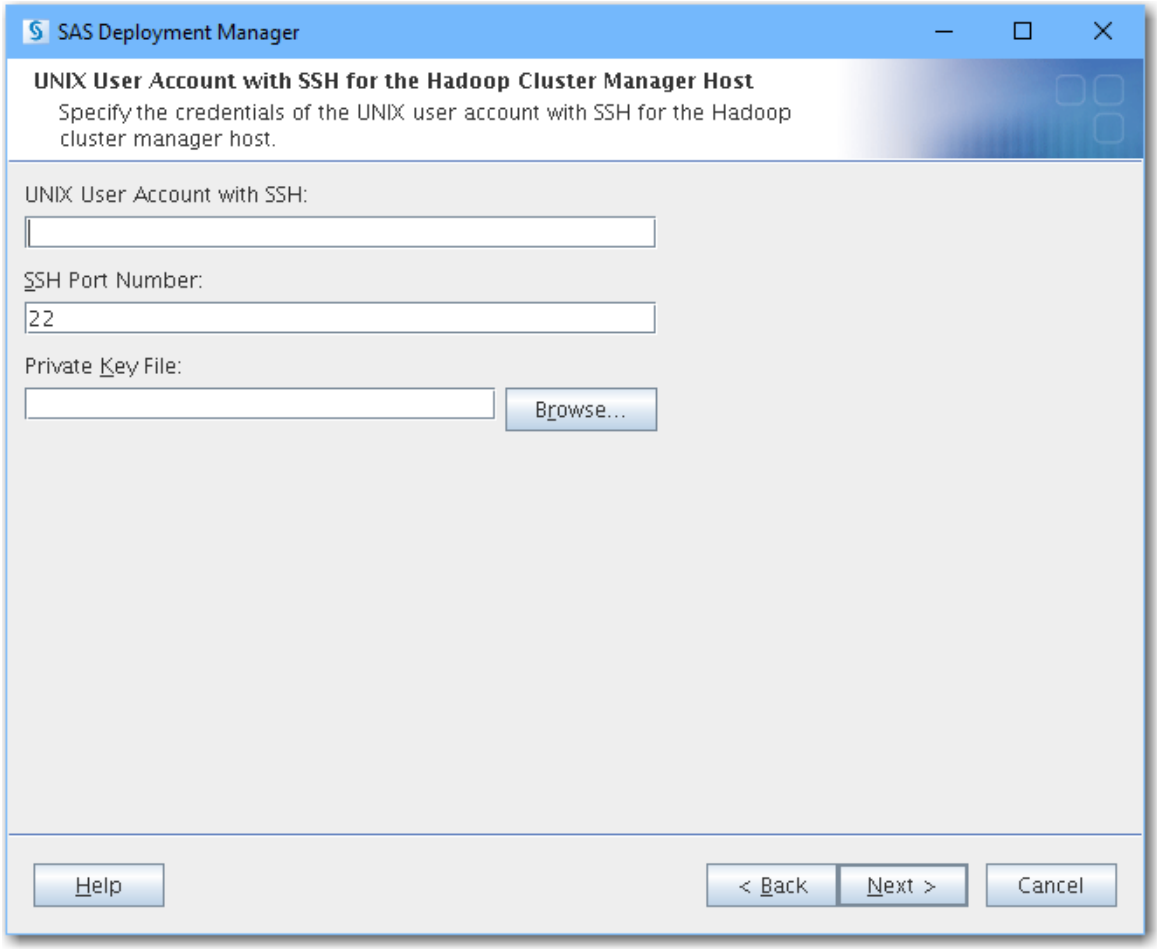
   The **Cluster Manager Host UNIX Account with SSH Information** page opens.



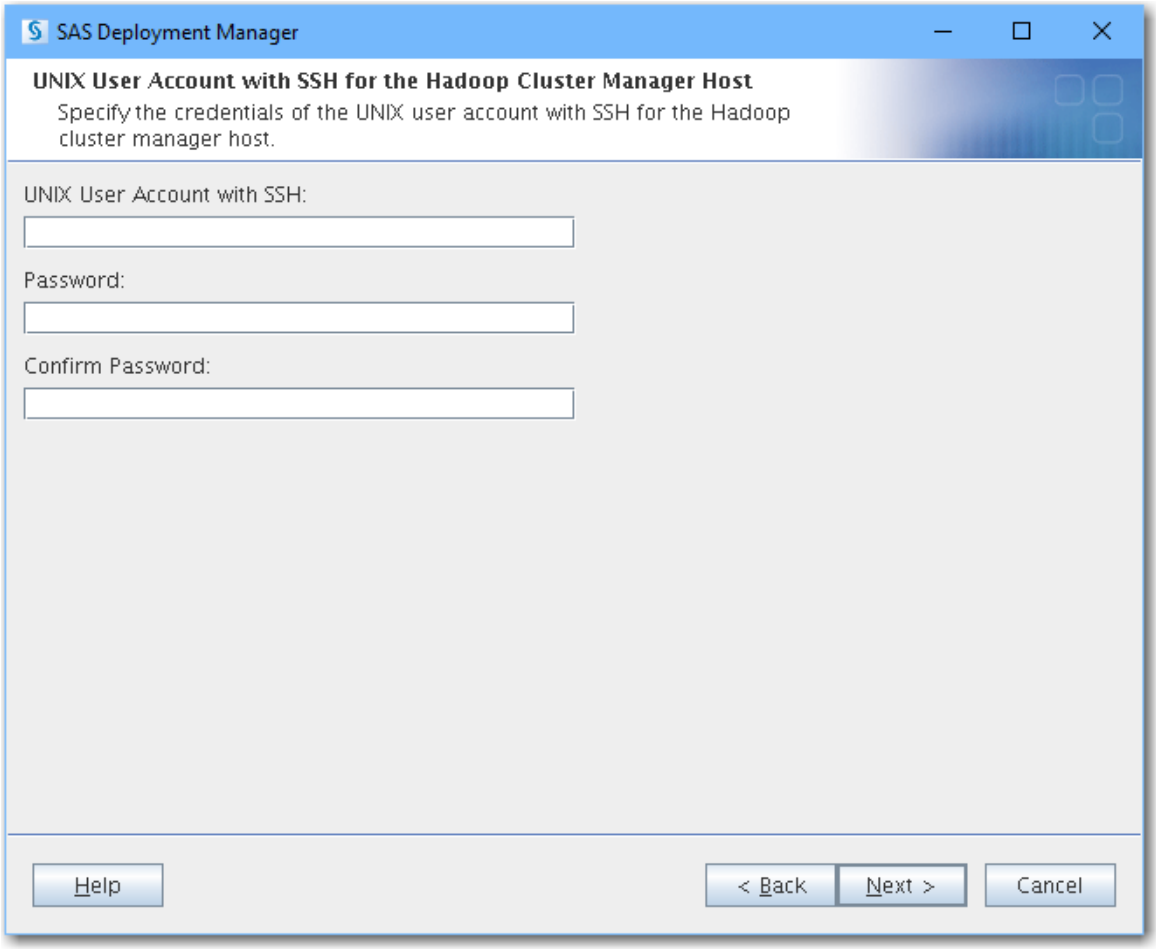
7. Specify whether you want to use a password or a private key file of the UNIX user account with SSH for the cluster manager host.

   Click **Next**.

   If you choose Yes, the **UNIX User Account with SSH for the Hadoop Cluster Manager Host** page opens that prompts you to browse for the private key file.



If you choose No, the **UNIX User Account with SSH for the Hadoop Cluster Manager Host** page opens and prompts you for a password.



Continue with Step 8 on page 15.

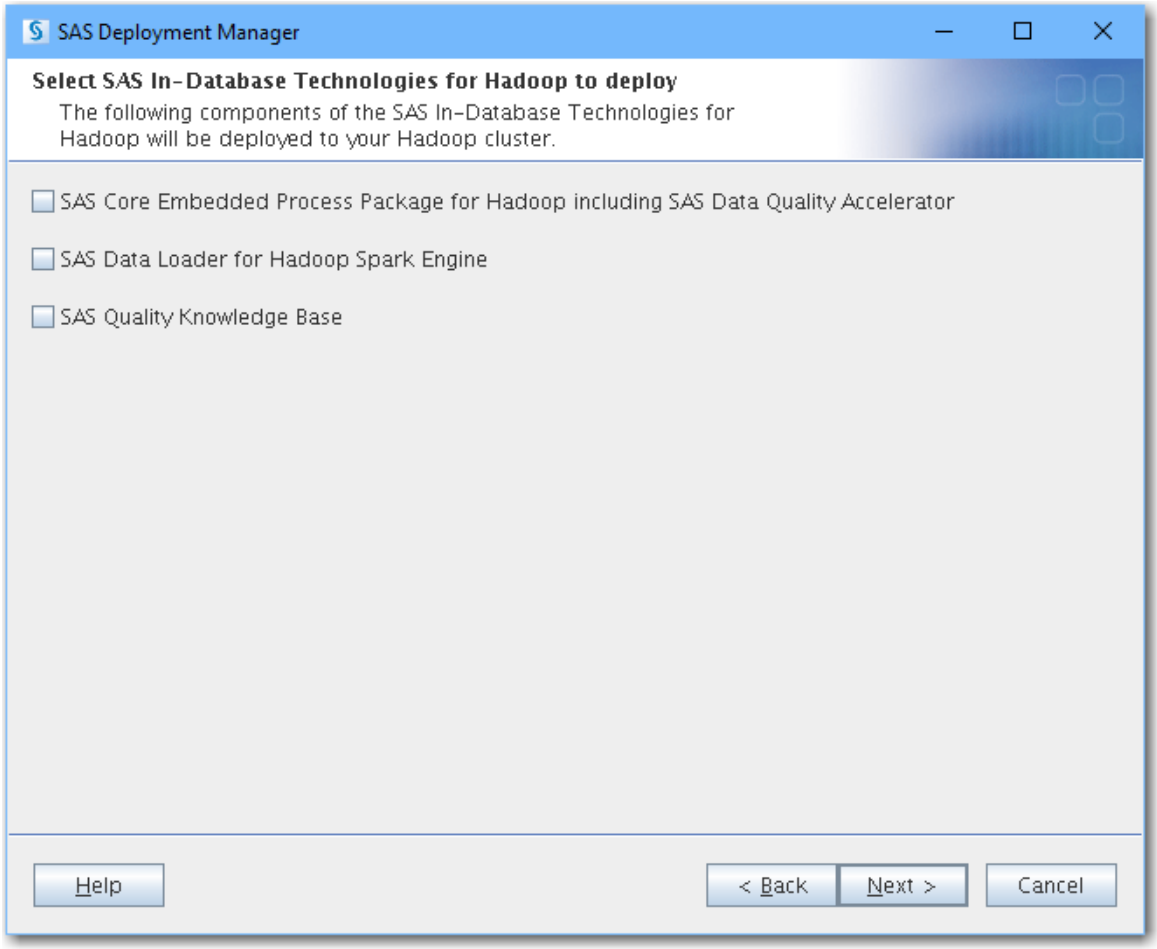
8. Perform one of the following actions depending on whether you chose to enter a password or provide a private key file:

   - If you chose No on the **Cluster Manager Host UNIX Account with SSH Information** panel, enter the root SSH account that has access to the cluster manager and password or enter a non-root SSH account if that account can execute sudo without entering a password.

     Note: For Cloudera, the SSH account must have Write permission to the `/opt/cloudera` directory. Otherwise, the deployment completes with errors.

   - If you chose Yes on the **Cluster Manager Host UNIX Account with SSH Information** panel, click **Browse** and navigate to the location of the private key file.

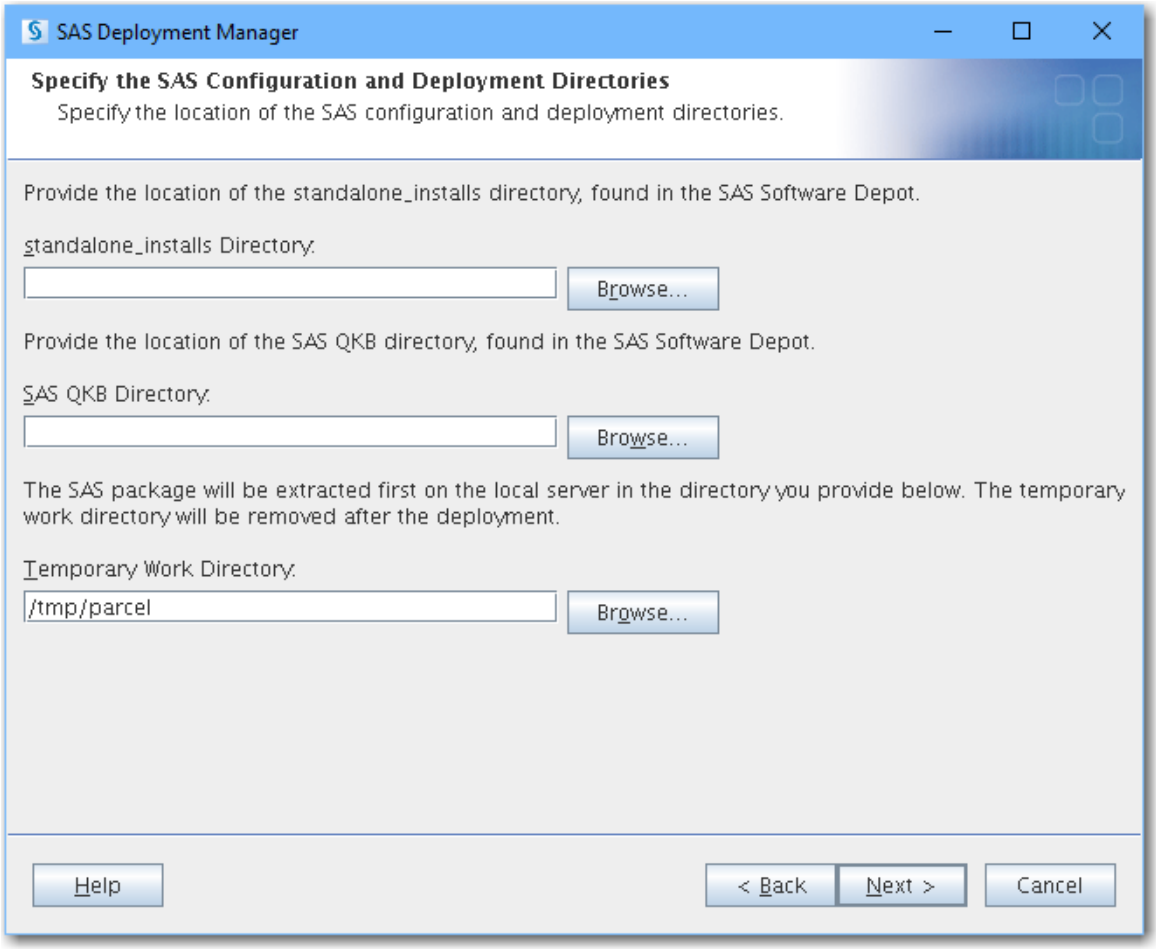
   Click **Next**. The **Select SAS In-Database Technologies for Hadoop to deploy** page opens.



9. Select which components to deploy on the Hadoop cluster.

   You can choose one, two, or all three components. However, the SAS Data Quality Accelerator, the SAS Data Loader for Hadoop Spark Engine, and the SAS Quality Knowledge Base (QKB) are available only if you license SAS Data Loader for Hadoop.

   Click **Next**. A page similar to the **Specify the SAS Configuration and Deployment Directories** page opens. What actually displays on this page depends on what components you choose to deploy.



10. Enter the location of the SAS configuration and deployment directories:

    a. Enter (or navigate to) the location of the **`/standalone_installs`** directory. This directory was created when your SAS Software Depot was created by the SAS Download Manager.

    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.

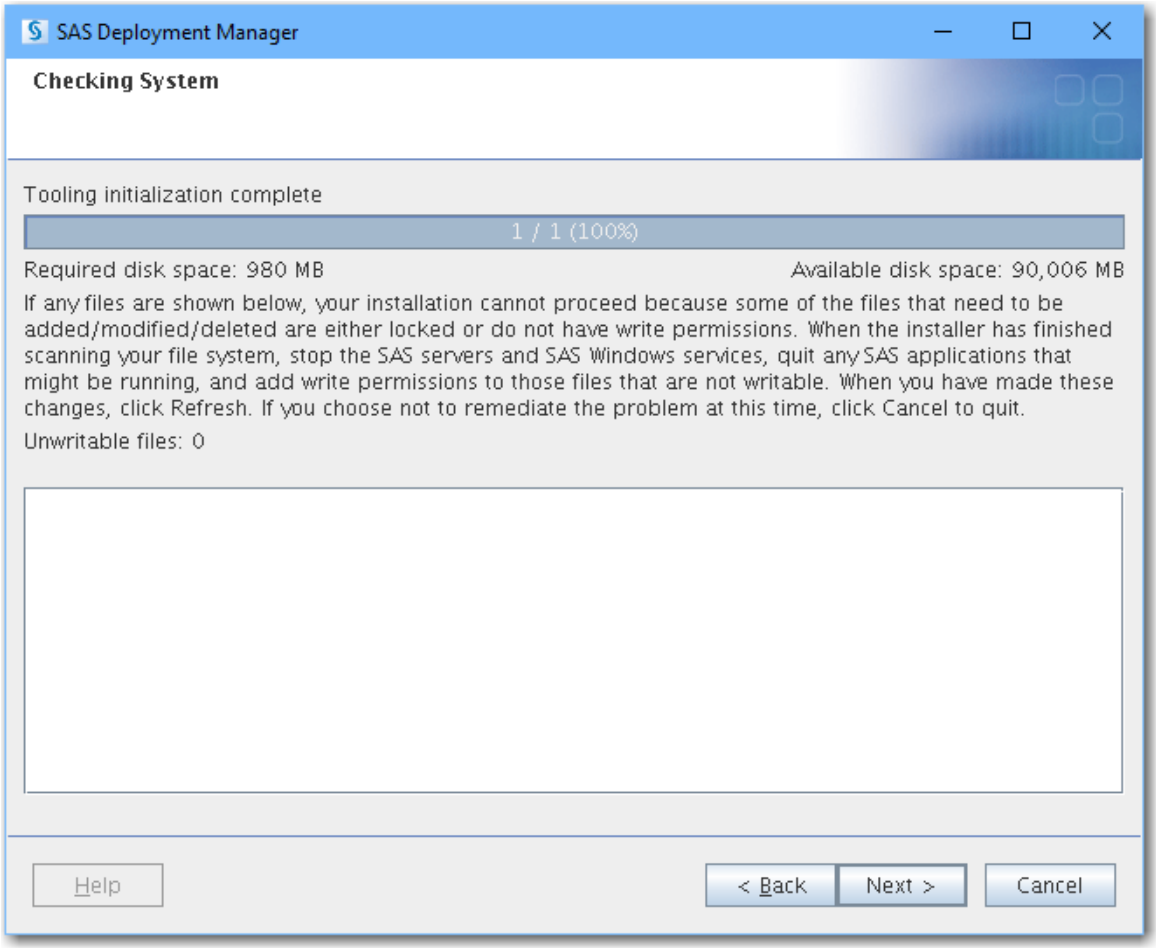
    b. (Optional) If you choose to deploy the SAS Quality Knowledge Base, enter (or navigate to) the location of the QKB directory, which can be found in the SAS Software Depot.

    Note: You obtain the SAS Quality Knowledge Base from your SAS Administrator. The SAS administrator identifies the SAS Quality Knowledge Base that SAS Data Loader for Hadoop should use for your site. For more information, see *SAS Data Loader for Hadoop: Installation and Configuration Guide*.

Note: A SAS Quality Knowledge Base must be deployed to enable data cleansing directives in SAS Data Loader for Hadoop.

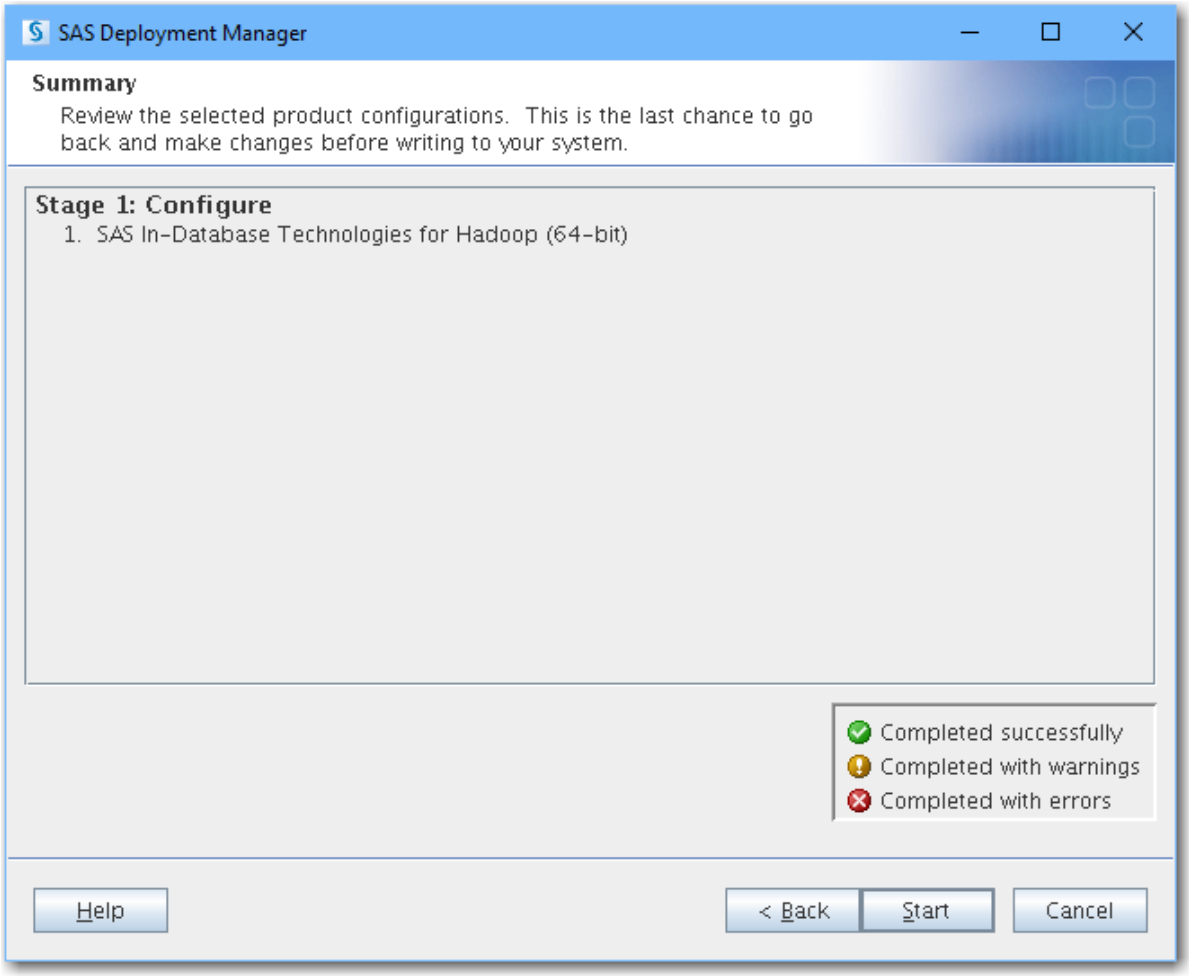
c. Enter (or navigate to) a temporary work directory on the local server where the package or stack is placed. The working directory is removed when the deployment is complete.

Click **Next**. The **Checking System** page opens, and a check for locked files and Write permissions is performed.



11. 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.



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

    Note: It takes time to complete the configuration. If your cluster is secured with Kerberos, it could take longer.

    Note: The product that appears on this page is the SAS product that is associated with the in-database deployment package for Hadoop. This package includes the SAS Embedded Process and possibly other components. Note that a separate license might be required to use the SAS Embedded Process and these other components.

    If the configuration is successful, the page title changes to **Deployment Complete** and a green check mark is displayed beside **SAS In-Database Technologies for Hadoop (64-bit)**.

    Note: Part of the configuration process runs SAS code to validate the environment. A green check mark indicates that the SAS Deployment Manager was able to create the SAS Embedded Process parcel or stack and then verify that the parcel or stack was copied to the cluster manager node.

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

13. Click **Next** to close the SAS Deployment Manager.

    A log file is written to the `%HOME/.SASAppData/SASDeploymentWizard` directory on the client machine.

14. Continue the installation process.

For more information, see "Deploying the SAS Embedded Process Parcel on Cloudera" on page 20 or "Deploying the SAS Embedded Process Stack on Hortonworks, IBM BigInsights, or Pivotal HD" on page 22.

Deploying the SAS Embedded Process Parcel on Cloudera

After you run the SAS Deployment Manager to create the SAS Embedded Process parcel, you must distribute and activate the parcel on the cluster. Follow these steps:

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: More than one SAS Embedded Process parcel can be deployed on your cluster, but only one parcel can be activated at one time. Before activating a new parcel, deactivate the old one.

Note: If you have licensed and downloaded SAS Data Loader for Hadoop or SAS Contextual Analysis In-Database Scoring for Hadoop, other SAS components are silently deployed at the same time as the SAS Embedded Process for Hadoop. Other configuration is required as noted in step 8. For more information about what components are also deployed, see "Overview of the In-Database Deployment Package for Hadoop" in *SAS In-Database Products: Administrator's Guide*.

1. Log on to Cloudera Manager.
2. Distribute the parcel to all nodes and create the SASEPHome directory.
   a. From the menu bar, choose **Hosts** ⇨ **Parcels**.

      The **SASEP** parcel is located under your cluster. An example name is p0.1.

      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 and create the SASEPHome directory.

      You can log on to the node and show the contents in the `/opt/cloudera/parcel` directory.
3. Click **Activate**.

   This step creates a symbolic link to the SAS Hadoop JAR file.

   When prompted, click **Close**.
4. Add the **SASEP** service and create the SAS Embedded Process configuration file in HDFS.
   a. Navigate to the 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** page appears.

c. Select the **SASEP** service and 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: The dependencies are automatically selected for this service.

e. On the **Add Service Wizard** ⇨ **Customize Role Assignments** page, select a node for the service.

Note: If your cluster is secured with Kerberos, in the next step, you must have a valid Kerberos ticket, so be sure to select a node that has a Kerberos ticket for the HDFS user on that node.

Choose any single node that is part of your cluster and where HDFS is a client.

Click **OK** and then click **Continue**. The **Add a SASEP to Cluster *cluster-name*** page appears.

f. Enter the name of the HDFS user. Click **Continue** and then click **Finish**.

Note: The default HDFS user name is *hdfs*. However, you can enter a custom HDFS user name.

Note: If your cluster is secured with Kerberos, the host that you select must have a valid ticket for the HDFS user.

The ep-config.xml file is created and added to the HDFS **`/sas/ep/config`** directory. This task is done in the host that you select.

g. After the SAS Embedded Process ep-config.xml file is created, Cloudera Manager starts the SAS Embedded Process service. This step is not required. MapReduce is the only service that is required for the SAS Embedded Process. You must stop the SAS Embedded Process service immediately when the task that adds the SAS Embedded Process is finished. The SAS Embedded Process service no longer needs to be stopped or started.

5. Verify that the ep-config.xml file exists in the **`/sas/ep/config`** directory of the host that you selected in step 4e.

6. Review any additional configuration that might be needed depending on your Hadoop distribution.

For more information, see "Additional Configuration for the SAS Embedded Process" in *SAS In-Database Products: Administrator's Guide*.

7. Validate the deployment of the SAS Embedded Process by running a program that uses the SAS Embedded Process and the MapReduce service. An example is a scoring program.

8. If you have licensed and downloaded the following SAS software, additional configuration is required:

- SAS Contextual Analysis In-Database Scoring for Hadoop

  For more information, see *SAS Contextual Analysis In-Database Scoring for Hadoop: Administrator's Guide*.

- SAS Data Loader for Hadoop

  For more information, see *SAS Data Loader for Hadoop: Installation and Configuration Guide*.

- SAS High-Performance Analytics

For more information, see *SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide*.

Deploying the SAS Embedded Process Stack on Hortonworks, IBM BigInsights, or Pivotal HD

Deploying the SAS Embedded Process Stack for the First Time

After you run the SAS Deployment Manager to create the SAS Embedded Process stack, you must deploy the stack on the cluster. Follow these steps:

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 the SAS Embedded Process stack already exists on your cluster, follow the instructions in "Deploying a New Version of the SAS Embedded Process Stack" on page 24.

Note: If you have licensed and downloaded SAS Data Loader for Hadoop or SAS Contextual Analysis In-Database Scoring for Hadoop, other SAS components are silently deployed at the same time as the SAS Embedded Process for Hadoop. Other configuration is required as noted in step 14. For more information about what components are also deployed, see "Overview of the In-Database Deployment Package for Hadoop" in *SAS In-Database Products: Administrator's Guide*.

1. Log on to the machine that is hosting Ambari.
2. Start the Ambari server and log on.
3. If the requiretty option was enabled when you deployed the SAS Embedded Process, you must restart the Ambari server at this time. Otherwise, skip to step 4.
   a. Log on to the cluster.
      ```
      sudo - su
      ```
   b. Restart the Ambari server.
      ```
      ambari-server restart
      ```
   c. Start the Ambari server and log on.
4. Click **Actions** and choose **+ Add Service**.
   The **Add Service Wizard** page appears.
5. Select **Choose Services**.
   The **Choose Services** panel appears.
6. In the **Choose Services** panel, select **SASEP**. Click **Next**.
   The **Assign Slaves and Clients** panel appears.

7. In the **Assign Slaves and Clients** panel, select items under **Client** where you want the stack to be deployed.

   Note: You should always select NAMENODE as one of the clients and NAMENODE should have these two client components installed: HDFS_CLIENT and HCAT_CLIENT.

   The **Customize Services** panel appears.

   The SASEP stack is listed under **activated_version**. An example name is s0.1.

8. Do not change any settings on the **Customize Services** panel.

   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.

   Click **Next**. The **Review** panel appears.

9. Review the information about the panel. If everything is correct, click **Deploy**.

   The **Install, Start, and Test** panel appears. After the SAS Embedded Process stack is installed on all nodes, click **Next**.

   The **Summary** panel appears.

10. Click **Complete**. The SAS Embedded Process stack is now installed on all nodes of the cluster.

    The **SASEP** service is displayed on the Ambari dashboard.

11. Verify that the SAS Embedded Process configuration file, ep-config.xml, exists in the `/sas/ep/config` directory.

12. Review any additional configuration that might be needed depending on your Hadoop distribution.

    For more information, see "Additional Configuration for the SAS Embedded Process" in *SAS In-Database Products: Administrator's Guide*.

13. Validate the deployment of the SAS Embedded Process by running a program that uses the SAS Embedded Process and the MapReduce service. An example is a scoring program.

14. If you have licensed and downloaded the following SAS software, additional configuration is required:

    - SAS Contextual Analysis In-Database Scoring for Hadoop

      For more information, see *SAS Contextual Analysis In-Database Scoring for Hadoop: Administrator's Guide*.

    - SAS Data Loader for Hadoop

      For more information, see *SAS Data Loader for Hadoop: Installation and Configuration Guide*.

    - SAS High-Performance Analytics

      For more information, see *SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide*.

Deploying a New Version of the SAS Embedded Process Stack

More than one SAS Embedded Process stack can be deployed on your cluster, but only one stack can be activated at one time. After you run the SAS Deployment Manager to create the SAS Embedded Process stack, follow these steps to deploy an additional SAS Embedded Process stack when one already exists on your cluster.

1. Log on to the machine that is hosting Ambari.
2. Restart the Ambari server and log on to the Ambari manager.
3. Select **SASEP**.

   In the **Services** panel, a restart symbol appears next to **SASEP**. The **Configs** tab indicates that a restart is required.
4. Click **Restart**.
5. Click **Restart All**.

   After the service is restarted, the previous version of the SAS Embedded Process still appears in the **activated_version** text box on the **Configs** tab. All deployed versions of the SAS Embedded Process stack should appear in the **sasep_allversions** text box.
6. Refresh the browser.

   The new version of the SAS Embedded Process should now appear as the **activated_version** text box on the **Configs** tab.

If, at any time, you want to activate another version of the SAS Embedded Process stack, follow these steps:

1. Enter the version number in the **activated_version** text box on the **Configs** tab.
2. Click **Save**.
3. Add a note describing your action (for example, “Changed from version s01.1 to s01.2”), and click **Next**.
4. Click **Restart**.
5. Click **Restart All**.
6. Refresh Ambari.

   The new service is activated.
7. Review any additional configuration that might be needed depending on your Hadoop distribution.

   For more information, see “Additional Configuration for the SAS Embedded Process” in *SAS In-Database Products: Administrator’s Guide*.
8. Validate the deployment of the SAS Embedded Process by running a program that uses the SAS Embedded Process and the MapReduce service. An example is a scoring program.
9. If you have licensed and downloaded the following SAS software, additional configuration is required:

   - SAS Contextual Analysis In-Database Scoring for Hadoop

     For more information, see *SAS Contextual Analysis In-Database Scoring for Hadoop: Administrator’s Guide*.
   - SAS Data Loader for Hadoop

For more information, see *SAS Data Loader for Hadoop: Installation and Configuration Guide*.

- SAS High-Performance Analytics

  For more information, see *SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide*.

Chapter 3

Deploying SAS Text Analytics Scoring Models

Overview of Model Deployment

To deploy a SAS Contextual Analysis text analytics (binary) model, follow these steps:

1. Obtain a text analytics model that includes score code for concept extraction (.li file), categorization (.mco file), or sentiment analysis (.sam).
2. Copy the score code model to the Hadoop master node (NameNode).
3. Use ta_push.sh to deploy the SAS Contextual Analysis score code model in the cluster.

Obtaining a Text Analytics Model

You can obtain the text analytics models from SAS Contextual Analysis. To generate the score code, see the section "Viewing and Downloading Code" in Chapter 2 of *SAS Contextual Analysis: User's Guide*.

- To locate the binary files for concepts and categories, see "Locating the Rules Files (LI and MCO)" in *SAS Contextual Analysis: Administrator's Guide*.

- To locate language-specific SAM (binary) files that are used in SAS Contextual Analysis, see the file **\tktg\sasmisc*nn*-base.sam** (Windows) or **misc/tktg/*nn*-base.sam** (UNIX) under your SAS installation directory. Note that the first two characters (*nn*) of the .sam filename denote the licensed language. For example, in **en-base.sam** , the **en** denotes English.
- To locate sentiment (SAM) files that you created using SAS Sentiment Analysis Studio, see *SAS Sentiment Analysis Studio 12.2: User's Guide*.

Copying the Text Analytics Model to the Hadoop NameNode

After you have obtained a text analytics model, you must copy it to the Hadoop NameNode. It is recommended that you copy the model to a temporary staging area, such as **/tmp/tastage**. You can copy the model to the Hadoop NameNode by using a file transfer command such as FTP or SCP, or by mounting the file system where the model is located on the Hadoop NameNode.

The following example shows how you might copy a model that exists on a Linux system to the Hadoop NameNode. The example uses the secure copy with the -r flag to recursively copy the specified directory. For the example, assume the following:

- The host name of the client desktop system where the model is installed is **desktop123**
- The location where the model is located in client desktop system.is **/opt/sas/ta/share**
- The host name of the Hadoop NameNode is **hmaster456**
- The target location on the NameNode is **/tmp/tastage**

To copy the model from the client desktop to the NameNode, issue this command:

```
scp -r /opt/sas/ta/share hmaster456:/tmp/tastage
```

About the ta_push.sh Executable File

SAS Contextual Analysis In-Database Scoring for Hadoop provides the ta_push.sh executable file to enable you to deploy the SAS text analytics models on Hadoop cluster nodes. The ta_push.sh file copies the specified model to a user-specified location on each of the Hadoop nodes. The ta_push.sh file automatically discovers all nodes in the cluster and deploys the model to the specified target model path on each of the cluster nodes.

Executing ta_push.sh

The ta_push.sh file must be run as the root user. The root user becomes the HDFS user in order to detect the nodes in the cluster.

Execute ta_push.sh as follows:

cd EPInstallDir/SASEPHome/bin

./ta_push.sh -s *source_path* **-t** *target_path*

where *source_path* specifies the path on the NameNode where you copied your model, and *target_path* specifies the path on each of the Hadoop nodes where you deployed your model.

Here is an example of the command:

```
./ta_push.sh -s /tmp/tastage/en-ne.li -t /opt/sas/ta/model/en-ne.li
```

Troubleshooting the Text Analytics Model Deployment

The text analytics model deployment can fail for the following reasons:

Problem: You did not obtain a Kerberos ticket before attempting to run ta_push.sh in a Kerberos environment.
Solution: Obtain the ticket and rerun.

Problem: You executed ta_push.sh from a directory other than the SASEPHome directory.
Solution: Run the ta_push.sh script from **EPInstallDir/SASEPHome/bin**.

Problem: There is insufficient space in the **/tmp** directory for ta_push.sh to run.
Solution: Clear space and try again.

ta_push.sh: Reference

Overview

The ta_push.sh file is created in the **EPInstallDir/SASEPHome/bin** directory by the SAS In-Database technology install script. You must execute ta_push.sh from this directory.

By default, ta_push.sh automatically discovers all nodes in the cluster and deploys the text analytics model to the specified target model path on each cluster nodes. Flags are provided to enable you deploy the model to specific nodes or a group of nodes. If you are expanding your Hadoop cluster by adding new nodes after the initial deployment, you might want to use one of these flags to deploy the model to these nodes and avoid redeploying to the entire cluster.

Run ta_push.sh as the root user. The root user becomes the HDFS in order to detect the nodes in the cluster.

Note: Only one model can be deployed to each Hadoop node at a time.

Syntax

./ta_push.sh -s source_path -t target_path

Required Arguments

-s source_path
specifies the pathname of the source model on the NameNode.

-t target_path
specifies the pathname of the target model on each of the Hadoop nodes.

Options

-h hostname
specifies the host name of the computers or computers on which to perform the deployment.

-f hostfile
specifies the name of a file that contains a list of host names on which to perform the deployment.

-?
displays usage information

-l logfile
directs status information to the specified log file instead of directing to standard output.

-r
Removes the model from Hadoop nodes.

-v
specifies verbose output.

Examples

To deploy models to one or more nodes that are specified on a command line, execute the following command:

```
./ta_push.sh -h hostname1 [-h hostname2] -s source_path -t
target_path
```

To deploy models using a file that contains a list of node names, execute the following command:

```
./ta_push.sh -f hostfile -s source_path -t target_path
```

To remove a model from the Hadoop nodes, execute the following command:

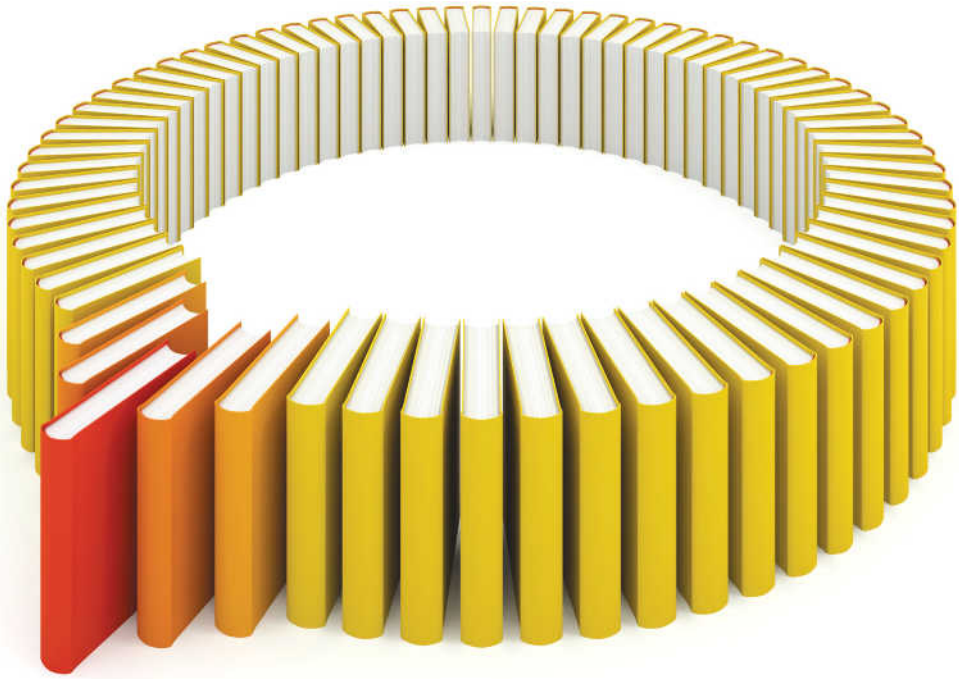
```
./ta_push.sh -r target_path
```

Recommended Reading

- *SAS Contextual Analysis: Administrator's Guide*
- *SAS Contextual Analysis: User's Guide*
- *SAS Contextual Analysis In-Database Scoring for Hadoop: User's Guide*
- *SAS In-Database Products: Administrator's Guide*
- *SAS In-Database Products: User's Guide*
- SAS Global Forum Paper, "Exploring SAS Embedded Process Technologies on Hadoop"

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.