THE POWER TO KNOW.

SAS® Data Loader 2.4 for Hadoop

Cluster Deployment Guide

Cloudera Manager and Ambari

SAS® Documentation

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2016. *SAS® Data Loader 2.4 for Hadoop: Cluster Deployment Guide (Cloudera Manager and Ambari)*. Cary, NC: SAS Institute Inc.

SAS® Data Loader 2.4 for Hadoop: Cluster Deployment Guide (Cloudera Manager and Ambari)

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513-2414.

February 2016

Contents

Chapter 1

Introduction

About this Guide

This guide is intended for customers who are installing SAS Data Loader for Hadoop, Cloudera or SAS Data Loader for Hadoop, Hortonworks. These products offer a streamlined installation process. Look for one of these product names in the software order email from SAS.

If your software order email uses only the term SAS Data Loader for Hadoop, please see the SAS Data Loader for Hadoop Administrator's Guide section of the *SAS In-Database Products: Administrator's Guide*.

SAS Data Loader and SAS In-Database Technologies for Hadoop

About SAS In-Database Technologies for Hadoop

SAS In-Database Technologies for Hadoop supports the Hadoop operations of SAS Data Loader for Hadoop. SAS Data Loader for Hadoop is web-client software that is installed as a vApp and is run on a virtual machine. The following products are included in SAS In-Database Technologies for Hadoop: SAS In-Database Deployment Package, SAS

Data Quality Accelerator, SAS Quality Knowledge Base, and SAS Data Management Accelerator for Spark.

SAS In-Database Deployment Package

The SAS In-Database Deployment Package includes the SAS Embedded Process and the SAS Hadoop MapReduce JAR files. The SAS Embedded Process runs within MapReduce to read and write data. You must deploy the SAS In-Database Deployment Package. Deploying and configuring the SAS In-Database Deployment Package needs to be done only once for each Hadoop cluster.

SAS Data Quality Accelerator and SAS Quality Knowledge Base

The data quality directives in SAS Data Loader for Hadoop are supported by SAS Data Quality Accelerator and the SAS Quality Knowledge Base (QKB). Both are required components for SAS Data Loader for Hadoop and are included in SAS In-Database Technologies for Hadoop. The QKB is a collection of files that store data and logic to support data management operations. A QKB is specific to a locale, that is, to a country and language. SAS Data Loader for Hadoop data quality directives reference the QKB when performing data quality operations on your data. It is recommended that you periodically update the QKB. For more information, see Chapter 6, “Updating the SAS Quality Knowledge Base (QKB),” on page 33.

SAS Data Management Accelerator for Spark

Spark is a processing engine that is compatible with Hadoop data. SAS Data Management Accelerator for Spark runs data integration and data quality tasks in a Spark environment. These tasks include mapping columns, summarizing columns, performing data quality tasks such as clustering and survivorship, and standardization of data. Deploy SAS Data Management Accelerator for Spark only if Spark is available on the cluster.

System Requirements

You can review the system requirements for the SAS Data Loader offering at the following location:

http://support.sas.com/documentation/installcenter/en/ikdmddhadpvofrsr/68979/PDF/default/sreq.pdf

Privileges

The Hadoop administrator installing SAS In-Database Technologies for Hadoop must have **`sudo`** or root privileges on the Hadoop cluster.

Support for the vApp User

You must configure the Hadoop cluster and provide certain values to the vApp user. For specific information about what you must provide, see "Providing vApp User Configuration Information" on page 26 and "Providing vApp User Configuration Information" on page 31.

Deployment

For detailed information about deployment, see Chapter 2, "Deployment in the Cluster," on page 5.

Chapter 2

Deployment in the Cluster

Getting Started with the Deployment

About the Deployment

This chapter describes deployment of SAS Data Loader 2.4 for Hadoop Cloudera and SAS Data Loader 2.4 for Hadoop Hortonworks. SAS sends an email to a contact person at your business or organization. This email specifies whether the product is SAS Data Loader 2.4 for Hadoop Cloudera or SAS Data Loader 2.4 for Hadoop Hortonworks and includes instructions for downloading a ZIP file. The ZIP file contains all product files that are required for installation of SAS In-Database Technologies for Hadoop on the Hadoop cluster. The contact person is responsible for making the ZIP file available to you.

Note: For further specific information about SAS In-Database Technologies for Hadoop, see Chapter 1, “Introduction,” on page 1.

Before Deploying

If you are installing a new version or reinstalling a previous version of SAS In-Database Technologies for Hadoop, you must deactivate or remove other existing SAS In-Database Technologies for Hadoop parcels or stacks after installing the new one. More than one parcel or stack can be deployed on your cluster, but only one parcel can be activated at a time. See "Deactivating or Removing Existing Versions" on page 13.

Overview of Deployment Steps

1. Review the Hadoop Environment topic in the system requirements for SAS Data Loader 2.4.
2. Configure Kerberos, if appropriate, and then provide required configuration values to the vApp user. See Chapter 5, "Configuring Kerberos," on page 27.
3. Identify a Windows server in a shared network location that is accessible to vApp users.
4. Obtain the ZIP file and extract zipped files. See "Obtain and Extract Zipped Files" on page 6.
5. Deploy services using Cloudera Manager or Ambari. See "Deploy Files" on page 7.
6. Edit the Hadoop configuration file. See "Edit SAS Hadoop Configuration Properties File" on page 11.
7. Collect required files from the Hadoop cluster. See "Collect Files" on page 12.
8. Make the required vApp directory available on the Windows server in the shared network location.
9. Configure the Hadoop cluster, and then provide required configuration values to the vApp user. See "Configure the Hadoop Cluster" on page 12.

Note: If you switch to a different distribution of Hadoop after the initial installation of SAS In-Database Technologies for Hadoop, you must reinstall and reconfigure SAS In-Database Technologies for Hadoop on the new Hadoop cluster.

Configure Kerberos

If you are using Kerberos, you must have all valid tickets in place on the cluster. When deploying SAS In-Database Technologies for Hadoop, the HDFS user must have a valid ticket. See Chapter 5, "Configuring Kerberos," on page 27. Provide the necessary configuration values to the vApp user.

Obtain and Extract Zipped Files

Download the ZIP file to a Linux machine that is accessible to the NameNode of the Hadoop cluster. Extract the contents of the file. Depending on whether your Hadoop

distribution is Cloudera or Hortonworks, you see one of the following file structures under **\products*package_name*\Admin**.

Figure 2.1 *Cloudera Manager File Structure*

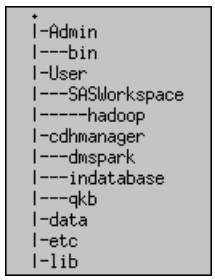
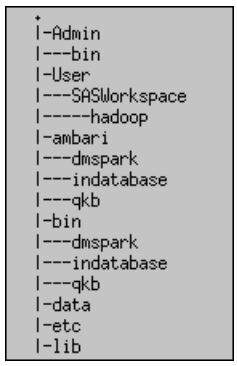


Figure 2.2 *Ambari File Structure*



Note: These examples illustrate a Windows environment, but the directories are the same under Linux.

Deploy Files

The following deployment steps assume that the extracted ZIP files are located directly on the Cloudera Manager or Ambari server or that you have placed the extracted ZIP files on a network location that is accessible to Cloudera Manager or the Ambari server.

Cloudera Manager

Creating Parcels

Navigate to the **Admin** directory and execute the following:

```
./bin/create_dl_parcel.sh -s pathname/Admin/cdhmanager -t pathname/Admin/parcels -v distro
```

where *pathname* is the location of the unzipped file structure and *distro* is one of the following Linux distributions: redhat5, redhat6, suse11x, ubuntu10, ubuntu12, ubuntu14, debian6, or debian7. You can also enter **-v** all to specify all distributions.

Deploying the Services to Cloudera

You must deploy SAS In-Database Deployment Package (SASEP) and SAS Quality Knowledge Base (SASQKB). Deploy SAS Data Management Accelerator for Spark

(SASDMSPARK) only if Spark is available on the cluster. It is recommended that you periodically update the QKB. For more information, see Chapter 6, "Updating the SAS Quality Knowledge Base (QKB)," on page 33.

1. Copy the following Custom Service Descriptor (CSD) files to the Cloudera Manager host, where *pathname* is the path to the unzipped files:

```
cp pathname/Admin/cdhmanager/indatabase/SASEP-9.43.jar /opt/cloudera/csd
cp pathname/Admin/cdhmanager/qkb/SASQKB-26.jar /opt/cloudera/csd
cp pathname/Admin/cdhmanager/dmspark/SASDMSPARK-2.4.jar /opt/cloudera/csd
```

2. Copy the following parcels to the Cloudera Manager host, where *pathname* is the path to the unzipped files:

```
cp pathname/Admin/cdhmanager/indatabase/SASEP-9.43.pdl.24-el6.parcel* /opt/cloudera/parcel-repo
cp pathname/Admin/cdhmanager/qkb/SASQKB-26.pdl.24-el6.parcel* /opt/cloudera/parcel-repo
cp pathname/Admin/cdhmanager/dmspark/SASDMSPARK-2.4.pdl.24-el6.parcel* /opt/cloudera/parcel-repo
```

3. On the Cloudera Manager host, change the ownership permissions on each of the following files:

```
chown cloudera-scm:cloudera-scm /opt/cloudera/csd/SASEP-9.43.jar
chown cloudera-scm:cloudera-scm /opt/cloudera/csd/SASQKB-26.jar
chown cloudera-scm:cloudera-scm /opt/cloudera/csd/SASDMSPARK-2.4.jar
chown cloudera-scm:cloudera-scm /opt/cloudera/parcel-repo/SASEP-9.43.pdl.24-el6.parcel*
chown cloudera-scm:cloudera-scm /opt/cloudera/parcel-repo/SASQKB-26.pdl.24-el6.parcel*
chown cloudera-scm:cloudera-scm /opt/cloudera/parcel-repo/SASDMSPARK-2.4.pdl.24-el6.parcel*
```

Note: If installing all three services, you can condense these commands as follows:

```
chown cloudera-scm:cloudera-scm /opt/cloudera/csd/SAS*
chown cloudera-scm:cloudera-scm /opt/cloudera/parcel-repo/SAS*
```

4. Restart the Cloudera Manager server by running the following command:

```
sudo service cloudera-scm-server restart
```

5. Log on to Cloudera Manager.
6. Activate each of the three parcels.

   Note: The following steps are iterative. For example, you must activate SASEP before activating SASQKB before activating SASDMSPARK.

   Select **Hosts** ⇨ **Parcels**. The parcels are located under your cluster. Complete the following for each parcel:

   a. Click **Distribute** to copy the parcel to all nodes.

   b. Click **Activate**. You are prompted either to restart the cluster or close the window.

   c. When prompted, click **Close**.

   CAUTION:
   Do not restart the cluster.

7. Add each of the three services. This creates files in HDFS.

   Note:

   - The following steps are iterative. For example, you must add the SASEP service before adding SASQKB before adding SASDMSPARK.
   - After adding a service, do not proceed to add another service without stopping the service that you have just added. If you proceed to add another

service while any of the other services are running, an error might be returned.

Complete the following for each service:

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 service and click **Continue**.

d. Select the dependencies for the service in the **Add Service Wizard** ⇨ **Select the set of dependencies for your new service** page. Click **Continue**.

Note: The dependencies are automatically selected for this service.

e. Select a node for the service in the **Add Service Wizard** ⇨ **Customize Role Assignments** page. Click **OK**, and then click **Continue**.

A file is added to HDFS for each of the services as follows:

- SASEP: **`/sas/ep/config/ep-config.xml`**
- SASQKB: **`/sas/qkb/default.idx`**
- SASDMSPARK: **`/sas/ep/config/dmp-config.xml`**

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

Note: If the services that you have just deployed are started, navigate to Cloudera Manager Home and stop them.

Ambari

Deploying the Services to Hortonworks

You must deploy SAS In-Database Deployment Package (SASEP) and SAS Quality Knowledge Base (SASQKB). Deploy SAS Data Management Accelerator for Spark (SASDMSPARK) only if Spark is available on the cluster. It is recommended that you periodically update the QKB. For more information, see Chapter 6, "Updating the SAS Quality Knowledge Base (QKB)," on page 33.

Note: You must complete the following steps on the Ambari Server host as the root user or as a user with **`sudo`** access

1. Copy the following files to the Ambari host, where *pathname* is the path to the unzipped files:

```
cp pathname/Admin/ambari/indatabase/SASEPINSTALL.gz /var/lib/ambari-server/resources
cp pathname/Admin/ambari/qkb/QKBINSTALL.gz /var/lib/ambari-server/resources
cp pathname/Admin/ambari/dmspark/SASDMSPARKINSTALL.gz /var/lib/ambari-server/resources
```

2. Execute the following command:

```
cd /var/lib/ambari-server/resources
```

3. Extract the contents of the following files:

```
tar -xvf SASEPINSTALL.gz
tar -xvf SASDMSPARKINSTALL.gz
```

Note: You do not need to extract QKBINSTALL.gz. During the deployment process, this file is extracted to each of the nodes in the cluster.

4. Copy and extract the following files to the Ambari host, where *pathname* is the path to the unzipped files:

```
cp pathname/Admin/ambari/indatabase/stacks.gz /var/lib/ambari-server/resources/stacks/HDP/2.0.6/services
cd /var/lib/ambari-server/resources/stacks/HDP/2.0.6/services
tar -xvf stacks.gz
rm stacks.gz

cp pathname/Admin/ambari/qkb/stacks.gz /var/lib/ambari-server/resources/stacks/HDP/2.0.6/services
cd /var/lib/ambari-server/resources/stacks/HDP/2.0.6/services
tar -xvf stacks.gz
rm stacks.gz

cp pathname/Admin/ambari/dmspark/stacks.gz /var/lib/ambari-server/resources/stacks/HDP/2.0.6/services
cd /var/lib/ambari-server/resources/stacks/HDP/2.0.6/services
tar -xvf stacks.gz
rm stacks.gz
```

5. Restart the Ambari server by running the following command:

   ```
   sudo ambari-server restart
   ```

6. Log on to Ambari.
7. 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 **SASEP SERVICE**, **SAS QKB**, and **SASDMSPARK**. Click **Next**.

      The **Assign Slaves and Clients** panel appears.

   c. In the **Assign Slaves and Clients** panel, select the NameNode, HDFS_CLIENT, and HCAT_CLIENT under **Client** where you want the stack to be deployed.

      The **Customize Services** panel appears.

      The SASQKB, SASDMSPARK, and SASEP SERVICE stacks are 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.

      SASEP SERVICE, SASQKB, and SASDMSPARK are displayed on the Ambari dashboard.

g. After deploying all of the services, verify that the following files exist in the Hadoop file system::

   - SASEP: **/sas/ep/config/ep-config.xml**
   - SASQKB: **/sas/qkb/default.idx**
   - SASDMSPARK: **/sas/ep/config/dmp-config.xml**

Edit SAS Hadoop Configuration Properties File

In the unzipped file structure, you must edit the file **Admin/etc/sas_hadoop_config.properties** to supply certain information that cannot be obtained automatically. Optional settings also exist that you might want to enable.

For the following section:

```
hadoop.client.config.filepath=<replace with full path>/User/SASWorkspace/hadoop/conf
hadoop.client.jar.filepath=<replace with full path>/User/SASWorkspace/hadoop/lib
hadoop.client.repository.path=<replace with full path>/User/SASWorkspace/hadoop/repository/
hadoop.client.configfile.repository=<replace with full path>/User/SASWorkspace/hadoop/repository
```

Replace <*replace with full path*> with the full path to the location where the ZIP file was unzipped.

For the following section:

```
hadoop.cluster.manager.hostname=
hadoop.cluster.manager.port=
hadoop.cluster.hivenode.admin.account=
hadoop.cluster.manager.admin.account=
```

Set **hadoop.cluster.manager.hostname** to the value of the host where either Cloudera Manager or Ambari is running.

Set **hadoop.cluster.manager.port** to the value of the port on which Cloudera Manager or Ambari is listening. Default values are provided.

Set **hadoop.cluster.hivenode.admin.account** to the value of a valid account on the machine on which the Hive2 service is running.

Set **hadoop.cluster.manager.admin.account** to the value of a valid Cloudera Manager or Ambari account.

For the following section:

```
hadoop.client.sasconfig.logfile.path=logs
hadoop.client.sasconfig.logfile.name=logs/sashadoopconfig/sashadoopconfig.log
hadoop.client.config.log.level=0
```

The default values of **logs** and **sashadoopconfig.log** create the directory **Admin/logs** and the filename **sashadoopconfig.log**, respectively. Both of these values can be changed if you prefer.

You can set the value of **hadoop.client.config.log.level** to 3 to increase the amount of information logged.

Note: If your distribution is secured with Kerberos,

- set **hadoop.cluster.hivenode.credential.type=kerberos**
- set **hadoop.client.config.log.level=3**

If you use Cloudera Manager and it manages multiple clusters, provide the name of the cluster to use for the value of **hadoop.cluster.manager.clustername=**.

Collect Files

Certain files must be collected from the Hadoop cluster and made available to the vApp user.

In the unzipped file structure, navigate to the **Admin** directory and run the following command:

```
./bin/hadoop_extract.sh
```

You are asked for two passwords:

- The cluster password is the password to the cluster manager administrative interface that corresponds to the hadoop.cluster.manager.admin.account name entered in the sas_hadoop_config.properties file.
- The hive password is the password for the SSH user account that is allowed to connect to the cluster that corresponds to the hadoop.cluster.hivenode.admin.account name entered in the sas_hadoop_config.properties file.

The script hadoop_extract.sh collects necessary files from the Hadoop cluster and stores them in two folders in the unzipped file structure:

***pathname*/User/SASWorkspace/hadoop/conf**

***pathname*/User/SASWorkspace/hadoop/lib**

Any collection issues are documented in the logs in ***pathname*/Admin/logs**. The script creates a backup of the original sas_hadoop_config.properties file.

Copy the complete **User** directory to a directory on a Windows server to which all vApp users have READ access. Inform all vApp users about the location of the **User** directory, which they must copy to their vApp client machines.

Configure the Hadoop Cluster

Complete configuration of the Hadoop cluster as described in Chapter 4, “Configuring the Hadoop Cluster,” on page 23. Provide the necessary configuration values to the vApp user.

Review any additional configuration that might be needed for the SAS Embedded Process, which is part of the In-Database Deployment Package. This is Hadoop distribution dependent. For more information, see Chapter 3, “Additional Configuration for the SAS Embedded Process,” on page 17.

Deactivating or Removing Existing Versions

About Deactivating and Removing

If you are installing a new version or reinstalling a previous version of SAS In-Database Technologies for Hadoop, you must deactivate or remove other existing parcels or stacks after installing the new one. You can have more than one parcel or stack for a particular product on the cluster, but only one can be active. At a minimum, you deactivate parcels or stacks that you do not want to use. You can also remove them from the cluster after deactivation.

Cloudera Manager

Example Names

Deactivation and removal of the parcels for SAS In-Database Deployment Package, SAS Quality Knowledge Base, and SAS Data Management Accelerator for Spark each follow the same procedure. The parcel names are SASEP, SASQKB, and SASDMSPARK, respectively. These names are represented in the following procedures by *parcel_name*. The configuration filenames for the SAS In-Database Deployment Package and SAS Data Management Accelerator for Spark are ep-config.xml and dmp-config.xml, respectively. The index filename for SAS Quality Knowledge Base is default.idx. These filenames are represented in the following procedures by *file_name*.

Deactivating

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

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/`**
      - The directory for SASQKB is **`/sas/qkb/`**
   c. Delete the *file_name* file.

Ambari

Example Names

Deactivation and removal of the stacks for SAS In-Database Deployment Package, SAS Quality Knowledge Base, and SAS Data Management Accelerator for Spark each follow the same procedure. The stack names are SASEP, SASQKB, and SASDMSPARK, respectively. These names are represented in the following procedures by *stack_service*. The configuration filenames for the SAS In-Database Deployment Package and SAS Data Management Accelerator for Spark are ep-config.xml and dmp-config.xml, respectively. The index filename for SAS Quality Knowledge Base is default.idx. These filenames are represented in the following procedures by *file_name*.

Deactivating

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. Optionally, add a note describing your action, and then click **Next**.

6. If you are deactivating more than one stack, finish all deactivation tasks before restarting the cluster.
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

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 **`indatabase`**, **`qkb`**, or **`dmspark`**. These directories are on the Linux machine where SAS In-Database Technologies for Hadoop 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.

   ```
   scp delete_stack.sh user@cluster-manager-host:/mytemp
   ```

3. Use this command to run the delete script.

   ```
   ./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 **SASEP SERVICE**.

6. Enter *y* to restart the Ambari server. The *stack_service* no longer appears.

Chapter 3

Additional Configuration for the SAS Embedded Process

Overview of Additional Configuration Tasks

After you have installed the SAS Embedded Process either manually or by using the SAS Deployment Manager, the following additional configuration tasks must be performed:

- “Additional Configuration Needed to Use HCatalog File Formats” on page 18.
- “Additional Configuration for Hortonworks 2.2” on page 19.
- “Adjusting the SAS Embedded Process Performance” on page 20.
- “Adding the SAS Embedded Process to Nodes after the Initial Deployment” on page 22.

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 standard MapReduce code and Pig. HCatalog is part of Apache Hive.

The SAS Embedded Process for Hadoop uses HCatalog to process the following complex, non-delimited file formats: Avro, ORC, Parquet, and RCFile.

Prerequisites for HCatalog Support

If you plan to access complex, non-delimited file types such as Avro or Parquet, you must perform these additional prerequisites:

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

File Type	Required Hive Version
Avro	0.14
ORC	0.11
Parquet	0.13
RCFile	0.6

SAS Server-Side Configuration

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. For Hadoop distributions that run with MapReduce 1, you must also manually add the HCatalog CLASSPATH to the MapReduce CLASSPATH.

Here is an example for a Cloudera distribution.

```
<property>
   <name>mapreduce.application.classpath</name>
   <value>/EPInstallDir/SASEPHome/jars/sas.hadoop.ep.apache205.jar,/EPInstallDir
   /SASEPHome/jars/sas.hadoop.ep.apache205.nls.jar,/opt/cloudera/parcels/
   CDH-5.2.0-1.cdh5.2.0.p0.36/bin/../lib/hive/lib/*,
   /opt/cloudera/parcels/CDH-5.2.0-1.cdh5.2.0.p0.36/lib/hive-hcatalog/libexec/
   ../share/hcatalog/*,/opt/cloudera/parcels/CDH-5.2.0-1.cdh5.2.0.p0.36/
   lib/hive-hcatalog/libexec/../share/hcatalog/storage-handlers/hbase/lib/*,
   $HADOOP_MAPRED_HOME/*,$HADOOP_MAPRED_HOME/lib/*,$MR2_CLASSPATH</value>
```

```
</property>
```

Here is an example for a Hortonworks distribution.

```
<property>
   <name>mapreduce.application.classpath</name>
   <value>/EPInstallDir/SASEPHome/jars/sas.hadoop.ep.apache205.jar,/SASEPHome/
   jars/sas.hadoop.ep.apache205.nls.jar,/usr/lib/hive-hcatalog/libexec/
   ../share/hcatalog/*,/usr/lib/hive-hcatalog/libexec/../share/hcatalog/
   storage-handlers/hbase/lib/*,/usr/lib/hive/lib/*,$HADOOP_MAPRED_HOME/
   share/hadoop/mapreduce/*,$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/
   lib/*</value>
</property>
```

Additional Configuration for Cloudera 5.4 or IBM BigInsights 4.0

If you are using Cloudera 5.4 or IBM Big Insights 4.0 with HCatalog sources, you must add the HADOOP_HOME environment variable in the Windows environment. An example of the value for this option is **HADOOP_HOME=c:\hadoop**.

The directory contains a subdirectory named **bin** that must contain the winutils.exe for your distribution. Please contact your distribution vendor for a copy of the winutils.exe file.

Additional Configuration for Hortonworks 2.2

Note: If you used the SAS Deployment Manager to install the SAS Embedded Process, this configuration task is not necessary. It was completed using the SAS Deployment Manager.

If you are installing the SAS Embedded Process on Hortonworks 2.2, you must manually revise the following properties in the mapred-site.xml property file on the SAS client side. Otherwise, an error occurs when you submit a program to Hadoop.

Use the **hadoop version** command to determine the exact version number of your distribution to use in place of **${hdp.version}**. This example assumes that the current version is 2.2.0.0-2041.

mapreduce.application.framework.path
Change

```
/hdp/apps/${hdp.version}/mapreduce/mapreduce.tar.gz#mr-framework
```

to

```
/hdp/apps/2.2.0.0-2041/mapreduce/mapreduce.tar.gz#yarn
```

mapreduce.application.classpath
Change

```
$PWD/mr-framework/hadoop/share/hadoop/mapreduce/*:$PWD/mr-framework/
hadoop/share/hadoop/mapreduce/lib/*:$PWD/mr-framework/hadoop/share/
hadoop/common/*:$PWD/mr-framework/hadoop/share/hadoop/common/lib/*:$PWD/
mr-framework/hadoop/share/hadoop/yarn/*:$PWD/mr-framework/hadoop/share/hadoop/
yarn/lib/*:$PWD/mr-framework/hadoop/share/hadoop/hdfs/*:$PWD/mr-framework/
hadoop/share/hadoop/hdfs/lib/*:/usr/hdp/${hdp.version}/hadoop/lib/
hadoop-lzo-0.6.0.${hdp.version}.jar:/etc/hadoop/conf/secure
```

to

```
/usr/hdp/2.2.0.0-2041/hadoop-mapreduce/*:/usr/hdp/2.2.0.0-2041/hadoop-mapreduce/
lib/*:/usr/hdp/2.2.0.0-2041/hadoop/*:/usr/hdp/2.2.0.0-2041/hadoop/lib/
*:/usr/hdp/2.2.0.0-2041/hadoop-yarn/*:/usr/hdp/2.2.0.0-2041/hadoop-yarn/lib/
*:/usr/hdp/2.2.0.0-2041/hadoop-hdfs/*:/usr/hdp/2.2.0.0-2041/hadoop-hdfs/lib/
*:/usr/hdp/2.2.0.0-2041/hadoop/lib/hadoop-lzo-0.6.0.2.2.0.0-2041.jar:/etc/
hadoop/conf/secure
```

yarn.app.mapreduce.am.admin-command-opts
Change

```
-Dhdp.version=${hdp.version}
```

to

```
-Dhdp.version=2.2.0.0-2041
```

yarn.app.mapreduce.am.command-opts
Change

```
-Xmx410m -Dhdp.version=${hdp.version}
```

to

```
-Xmx410m -Dhdp.version=2.2.0.0-2041
```

Note: If you upgrade your Hortonworks distribution and the version changes, you need to make this update again.

Adjusting the SAS Embedded Process Performance

Overview of the ep-config.xml File

You can adjust how the SAS Embedded Process runs by changing properties in the ep-config.xml file.

The ep-config.xml file is created when you install the SAS Embedded Process. By default, the file is located in the **`/sas/ep/config/ep-config.xml`** directory.

You can change property values that enable you to perform the following tasks:

- change trace levels

  For more information, see "Changing the Trace Level" on page 21.

- specify the number of SAS Embedded Process MapReduce 1 tasks per node

  For more information, see "Specifying the Number of MapReduce Tasks" on page 21.

- specify the maximum amount of memory in bytes that the SAS Embedded Process is allowed to use

  For more information, see "Specifying the Amount of Memory That the SAS Embedded Process Uses" on page 21.

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).

```
<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: Tracing requires that an **/opt/SAS** directory to exist on every node of the cluster when the SAS Embedded Process is installed. If the folder does not exist or does not have Write permission, the SAS Embedded Process job fails.

Specifying the Number of MapReduce Tasks

You can specify the number of SAS Embedded Process MapReduce Tasks per node by changing the sas.ep.superreader.tasks.per.node property in the ep-config.xml file. The default number of tasks is 6.

```
<property>
   <name>sas.ep.superreader.tasks.per.node</name>
   <value>number-of-tasks</value>
</property>
```

Specifying the Amount of Memory That the SAS Embedded Process Uses

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. The default value is 2147483647 bytes.

```
<property>
   <name>sas.ep.max.memory</name>
```

```
   <value>number-of-bytes</value>
</property>
```

Note: This property is valid only for Hadoop distributions that are running MapReduce 1.

If your Hadoop distribution is running MapReduce 2, this value does not supersede the YARN maximum memory per task. Adjust the YARN container limit to change the amount of memory that the SAS Embedded Process is allowed to use.

Adding the SAS Embedded Process to Nodes after the Initial Deployment

After the initial deployment of the SAS Embedded Process, additional nodes might be added to your cluster or nodes might need to be replaced. In these instances, you can install the SAS Embedded Process on the new nodes.

Follow these steps:

1. 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 Kerberos ticket to access HDFS. This can be done with kinit.

2. Navigate to the **/sas/ep/config/** directory on HDFS.

3. Remove the ep-config.xml file from HDFS.

   ```
   cd /sas/ep/config/
   hadoop fs -rm ep-config.xml
   ```

4. Run the sasep-admin.sh script and specify the nodes on which you want to install the SAS Embedded Process.

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

Chapter 4

Configuring the Hadoop Cluster

Configuring Components on the Cluster

Overview

After deploying the in-database deployment package, you must configure several components and settings on the Hadoop cluster in order for SAS Data Loader for Hadoop to operate correctly. These components and settings are explained in the following topics:

- “SQOOP and OOZIE” on page 23
- “JDBC Drivers” on page 24
- “ Spark Bin Directory Required in the Hadoop PATH” on page 24
- “User IDs” on page 25
- “Configuration Values” on page 25

SQOOP and OOZIE

Your Hadoop cluster must be configured to use OOZIE scripts.

Note: Ensure that Oozie 4.0 or later is installed. You must add the following as entries in the list for the oozie.service.SchemaService.wf.ext.schemas property:

- sqoop-action-0.4.xsd
- hive-action-0.3.xsd
- oozie-workflow-0.4.xsd

- shell-action-0.3.xsd (for Spark submission)

JDBC Drivers

SAS Data Loader for Hadoop leverages the SQOOP and OOZIE components installed with the Hadoop cluster to move data to and from a DBMS. The SAS Data Loader for Hadoop vApp client also accesses databases directly using JDBC for the purpose of selecting either source or target schemas and tables to move.

You must install on the Hadoop cluster the JDBC driver or drivers required by the DBMSs that users need to access.

SAS Data Loader for Hadoop supports the Teradata and Oracle DBMSs directly. You can support additional databases by selecting **Other** in the **Type** option on the SAS Data Loader for Hadoop Database Configuration dialog box. For more information about the dialog box, see the *SAS Data Loader for Hadoop: User's Guide*.

For Teradata and Oracle, SAS recommends that you download the following JDBC files from the vendor site:

Table 4.1 *JDBC Files*

Database	Required Files
Oracle	ojdbc6.jar
Teradata	tdgssconfig.jar and terajdbc4.jar *Note:* You must also download the Teradata connector JAR file that is matched to your cluster distribution, if available.

The JDBC and connector JAR files must be located in the OOZIE shared libs directory in HDFS, not in **`/var/lib/sqoop`**. The correct path is available from the oozie.service.WorkflowAppService.system.libpath property.

The default directories in the Hadoop file system are as follows:

- Hortonworks Hadoop clusters: **`/user/oozie/share/lib/lib_`*`version`*`/sqoop`**
- Cloudera Hadoop clusters: **`/user/oozie/share/lib/sharelib`*`version`*`/sqoop`**

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

After JDBC drivers have been installed and configured along with SQOOP and OOZIE, you must refresh sharelib, as follows:

```
oozie admin -oozie oozie_url -sharelibupdate
```

SAS Data Loader for Hadoop users must also have the same version of the JDBC drivers on their client machines in the **`SASWorkspace\JDBCDrivers`** directory. Provide a copy of the JDBC drivers to SAS Data Loader for Hadoop users.

Spark Bin Directory Required in the Hadoop PATH

SAS Data Loader for Hadoop supports the Apache Spark cluster computing framework. Spark support requires the addition of the Spark bin directory to the PATH environment variable on each Hadoop node.

Most Hadoop distributions include the Spark bin directory in **`/usr/bin`**.

User IDs

Kerberos

If your installation uses Kerberos authentication, see Chapter 5, "Configuring Kerberos," on page 27.

UNIX User Accounts and Home Directories

You must create one or more user IDs and enable certain permissions for the SAS Data Loader for Hadoop vApp user.

To configure user IDs, follow these steps:

1. Choose one of the following options for user IDs:
   - Create one user ID that any vApp user can use for login.

     Note: Do not use the super user, which is typically hdfs.
   - Create an individual user ID for each vApp user.
   - Map the user ID to a user principal for clusters using Kerberos.
2. Create UNIX user IDs on all nodes of the cluster and assign them to a group.
3. Create a user home directory and Hadoop staging directory in HDFS. The user home directory is **`/user/myuser`**. The Hadoop staging directory is controlled by the setting yarn.app.mapreduce.am.staging-dir in mapred-site.xml and defaults to **`/user/myuser`**.
4. Change the permissions and owner of **`/user/myuser`** to match the UNIX user.

   Note: The user ID must have at least the following permissions:
   - Read,Write, and Delete permission for files in the HDFS directory (used for Oozie jobs)
   - Read, Write, and Delete permission for tables in Hive

Configuration Values

You must provide the vApp user with values for fields in the SAS Data Loader for Hadoop Configuration dialog box. For more information about the SAS Data Loader for Hadoop Configuration dialog box, see the *SAS Data Loader for Hadoop: vApp Deployment Guide*. The fields are as follows:

Host
specifies the full host name of the machine on the cluster running the HiveServer2 server.

Port
specifies the number of the HiveServer2 server port on your Hadoop cluster. For most distributions, the default is 10000.

User ID
specifies the Hadoop user account that you have created on your Hadoop cluster for each user or for all of the vApp users.

Note:

- For Cloudera and Hortonworks user IDs, see "UNIX User Accounts and Home Directories" on page 25.

Password
if your enterprise uses LDAP, you must supply the vApp user with the LDAP password. This field must be blank otherwise.

Oozie URL
specifies the Oozie base URL. The URL is the property oozie.base.url in the file oozie-site.xml. The URL is similar to the following example: http://*host_name*:*port_number*/oozie/.

Although the Oozie web UI at this URL does not have to be enabled for Data Loader to function, it is useful for monitoring and debugging Oozie jobs. Confirm that the Oozie Web UI is enabled before providing it to the vApp user. Consult your cluster documentation for more information.

Providing vApp User Configuration Information

The configuration components and information that the Hadoop administrator must supply to the vApp user are summarized in the following tables:

Table 4.2 *Configuration Components and Information*

Component	Location of Description
JDBC drivers	See "JDBC Drivers" on page 24.
User IDs	See "User IDs" on page 25.

The SAS Data Loader for Hadoop vApp that runs on the client machine contains both Settings and Configuration dialog boxes. For more information about these dialog boxes, see the *SAS Data Loader for Hadoop: vApp Deployment Guide*.

The Configuration dialog box contains certain fields for which you must provide values to the vApp user. These fields are as follows:

Table 4.3 *Configuration Fields*

Field	Location of Description
Host	See "Host" on page 25.
Port	See "Port" on page 25.
User ID	See "User ID" on page 25.
Password	See "Password" on page 26.
Oozie URL	See "Oozie URL" on page 26.

Chapter 5

Configuring Kerberos

About Kerberos on the Hadoop Cluster

If your enterprise uses Kerberos security, you must have all valid tickets in place on the cluster. When SAS In-Database Technologies for Hadoop is deployed, the HDFS user must have a valid ticket.

Note:

- For all Hadoop distributions, the default HDFS user is **`hdfs`**.
- If you set a maximum lifetime for Kerberos tickets, ensure that the person deploying SAS In-Database Technologies for Hadoop is aware of the expiration date of the ticket.

After configuring Kerberos, provide the necessary configuration values to the vApp user. See "Providing vApp User Configuration Information" on page 31.

Note: SAS Data Loader for Hadoop does not provide Kerberos validation. All configuration values must be entered correctly in the SAS Data Loader for Hadoop vApp or errors result during its operation.

Client Configuration

Certain configuration must take place on the client machine that hosts the vApp. For example, the hosts file on the client machine must be modified to include the host name that is used to access SAS Data Loader for Hadoop. This host name must be the same

host name that is used to generate keytabs for Kerberos, as described in "Kerberos Configuration" on page 28. For more information, see the *SAS Data Loader for Hadoop: vApp Deployment Guide*.

Kerberos Configuration

Overview

The Kerberos topology contains multiple tiers. They are configured to communicate with the Kerberos Key Distribution Center (KDC) to allow authentication to flow from the SAS Data Loader for Hadoop client machine through to the Hadoop cluster. When you log on to the client machine, the KDC issues a ticket granting ticket (TGT), which is time stamped. This TGT is used by the browser to issue a ticket to access SAS Data Loader for Hadoop.

Two different types of Kerberos systems are available: AD (Windows Active Directory) and MIT. You might have either a realm for only AD Kerberos or mixed AD and MIT realms. A realm for only AD Kerberos protects the client machine, the vApp virtual machine, and the Hadoop cluster all through the AD domain controller. A realm for only AD Kerberos is simpler because it requires less client configuration.

In a common configuration of mixed realms, AD Kerberos protects both the client machine and the vApp virtual machine, whereas MIT Kerberos protects only the Hadoop cluster. The mixed realms can be configured such that AD Kerberos protects only the client machine, whereas MIT Kerberos protects both the Hadoop cluster and the vApp virtual machine. Finally, it is possible to configure an all-MIT environment using the MIT Kerberos for Windows libraries to authenticate the client. Which realm configuration is in use determines how you must configure Kerberos.

vApp

Overview

You must generate a Service Principal Name (SPN) and Kerberos keytab for the host, SAS, and HTTP service instances.

The following SPNs must be created to allow ticket delegation, where *hostname* represents the host name that you have created and *KRBREALM* represents your Kerberos realm:

- host/*hostname@KRBREALM*.
- SAS/*hostname@KRBREALM*. This allows single sign-on from the middle tier to the SAS Object Spawner.
- HTTP/*hostname@KRBREALM*. This allows single sign-on with the tc Server and the SASLogon web application.

Protecting the vApp with MIT Kerberos

When protecting the vApp using MIT Kerberos, the client machine must be configured to acquire tickets for the vApp from the correct realm. For more information, see the *SAS Data Loader for Hadoop: vApp Deployment Guide*. You must provide the name of the KDC server to the person configuring the client machine.

On a machine that is configured to communicate with the MIT Kerberos realm, generate the three SPNs and corresponding keytabs. For example, if the fully qualified domain name is dltest1.vapps.zzz.com issue the following commands:

```
$ kadmin -p user2/admin -kt /home/user2/user2_admin.keytab
  kadmin: addprinc -randkey +ok_as_delegate host/dltest1.vapps.zzz.com
  kadmin: ktadd -k $hostname/host.dltest1.keytab host/dltest1.vapps.zzz.com
  kadmin: addprinc -randkey +ok_as_delegate SAS/dltest1.vapps.zzz.com
  kadmin: ktadd -k $hostname/SAS.dltest1.keytab SAS/dltest1.vapps.zzz.com
  kadmin: addprinc -randkey +ok_as_delegate HTTP/dltest1.vapps.zzz.com
  kadmin: ktadd -k $hostname/HTTP.dltest1.keytab HTTP/dltest1.vapps.zzz.com
```

Note: You must enable the **`ok_as_delegate`** flag to allow ticket delegation in the middle tier.

Protecting the vApp with AD Kerberos

To generate SPNs and keytabs in AD Kerberos on Windows Server 2012, you must have administrator access to the Windows domain and then follow these steps:

1. Create Managed Service Accounts:
   a. Launch the Server Manager on the domain controller:
   b. Select **Server Manager** ⇨ **Tools** ⇨ **Active Directory Users and Computers**.
   c. Select **<domain name>** ⇨ **Managed Service Accounts**.
   d. In the right pane, click **New** ⇨ **User**.
   e. In the **User logon name** field, enter **`host/`*`fully-qualified-hostname`***. For example, enter **`host/dltest1.vapps.zzz.com`**, and then click **Next**.
   f. Enter and confirm a password.
   g. If you are configuring a server with an operating system older than Windows 2000, change the logon name to **`HTTP/`*`simple-hostname`***. For example, enter **`host/dltest1`**.
   h. Deselect **User must change password at next logon** and the select **Password never expires**.
   i. Click **Finish**.
   j. Repeat the previous steps for the SAS and HTTP service accounts.
2. Create SPNs for each SPN user. At a command prompt on the domain controller, enter the following commands using a fully qualified host name and simple host name. For example, you might use dltest1.vapps.zzz.com and **`dltest1`**:

   ```
   > setspn -A host/dltest1.vapps.zzz.com host_dltest1
   > setspn -A SAS/dltest1.vapps.zzz.com SAS_dltest1
   > setspn -A HTTP/dltest1.vapps.zzz.com HTTP_dltest1
   ```

3. Authorize ticket delegation:
   a. Launch the Server Manager on the domain controller.
   b. Select **Server Manager** ⇨ **Tools** ⇨ **Active Directory Users and Computers**.
   c. Select **View** ⇨ **Advanced Features**.
   d. Select **host/<vapp> user**. Right-click, and then select **Properties**.
   e. Select the **Delegation** tab.

f. Select **Trust this user for delegation to any service (Kerberos only)**, and then click **Apply**.

g. Navigate to the **Attribute Editor** tab

h. On the **Attribute Editor** tab, locate the **msDS-KeyVersionNumber** attribute. Record this number. Click **OK**.

i. Repeat the previous steps to authorize ticket delegation for the SAS and HTTP users.

4. Create keytabs for each SPN. For UNIX, continue with this step. For Windows, skip to Step 5 on page 30.

   a. At a command prompt, use the ktutil utility to create keytabs. Enter the following commands using a fully qualified host name, the realm for your domain, the password that you created, and the msDS-KeyVersionNumber. In the following host SPN keytab example, **`dltest1.vapps.zzz.com`**, **`AD.ZZZ.COM`**, **`Psword`**, and **`-k 2 -e arcfour-hmac`** are used for these values:

```
ktutil
ktutil:  addent -password -p host/dltest1.vapps.zzz.com@AD.ZZZ.COM -k 2 -e arcfour-hmac
Psword for host/dltest1.vapps.zzz.com@AD.ZZZ.COM :
ktutil:  addent -password -p host/dltest1.vapps.zzz.com@AD.ZZZ.COM -k 2 -e aes128-cts-hmac-sha1-96
Psword for host/dltest1.host.zzz.com@AD.ZZZ.COM :
ktutil:  addent -password -p host/dltest1.vapps.zzz.com@AD.ZZZ.COM -k 2 -e aes256-cts-hmac-sha1-96
Psword for host/dltest1.vapps.zzz.com@AD.ZZZ.COM :
ktutil:  wkt host.dltest1.keytab
ktutil:  quit
```

   b. Repeat the previous steps to create the SAS and HTTP keytabs.

5. To create keytabs for each SPN on Windows, follow these steps:

   a. At a command prompt, use the ktpass utility to create keytabs. Enter the following commands using a fully qualified host name, the realm for your domain, and any password (it does not have to be the password that you created earlier). In the following host SPN keytab example, **`dltest1.vapps.zzz.com`**, **`AD.ZZZ.COM`**, and **`Psword`** are used for these values:

```
ktpass.exe -princ host/dltest1.vapps.zzz.com@AD.ZZZ.COM -mapUser user@fully.qualified.domain -pass "Psword"
         -pType KRB5_NT_PRINCIPAL -out dltest1-host.keytab -crypto All
```

   b. Repeat the previous steps to create the SAS and HTTP keytabs.

6. Provide the keytabs to the vApp user.

Hadoop

Overview

The Hadoop cluster must be configured for Kerberos according to the instructions provided for the specific distribution that you are using.

Ensure that the following setting is correct on your cluster:

```
* hive.server2.enable.doAs = true
```

Configure Kerberos Trusts

If the Kerberos environment includes users or services authenticated by a realm other than the default realm of the cluster, you must configure the cluster to interpret principals from the trusted realm. This is the case when the cluster is protected by MIT Kerberos and the client is protected by Active Directory.

Cloudera

When the cluster is protected by MIT Kerberos, add **`AD_DOMAIN_REALM`** to Trusted Kerberos Realms under the HDFS configuration.

Other Distributions

When the cluster is protected by MIT Kerberos, you must set the properties hadoop.security.auth_to_local and oozie.authentication.kerberos.name.rules as follows:

```
RULE:[1:$1@$0](.*@\QAD_DOMAIN_REALM\E$)s/@\QAD_DOMAIN_REALM\E$//
RULE:[2:$1@$0](.*@\QAD_DOMAIN_REALM\E$)s/@\QAD_DOMAIN_REALM\E$//
RULE:[1:$1@$0](.*@\QMIT_DOMAIN_REALM\E$)s/@\QMIT_DOMAIN_REALM\E$//
RULE:[2:$1@$0](.*@\QMIT_DOMAIN_REALME$)s/@\QMIT_DOMAIN_REALM\E$//
DEFAULT
```

An example of RULE 1 and RULE 2 for *AD_DOMAIN_REALM* is as follows:

```
RULE:[1:$1@$0](.*@\QDAFFY_KRB5.COM\E$)s/@\QDAFFY_KRB5.COM\E$//
RULE:[2:$1@$0](.*@\QDAFFY_KRB5.COM\E$)s/@\QDAFFY_KRB5.COM\E$//
DEFAULT
```

SAS LASR Analytic Server

Integration of SAS Data Loader for Hadoop with a SAS LASR Analytic Server is possible only in an AD Kerberos environment. SAS Data Loader for Hadoop cannot be integrated with SAS LASR Analytic Server in a mixed AD and MIT Kerberos environment.

A public key is created as part of SAS Data Loader for Hadoop vApp configuration and is placed in the SAS Data Loader for Hadoop shared folder. This public key must also exist on the SAS LASR Analytic Server grid. The public key must be appended to the authorized_keys file in the .ssh directory of that user.

For more information about the SAS LASR Analytic Server administrator, see “LASR Analytic Servers Panel” in the *SAS Data Loader for Hadoop: User’s Guide*.

Providing vApp User Configuration Information

The SAS Data Loader for Hadoop vApp that runs on the client machine contains a Settings dialog box in the SAS Data Loader: Information Center. For more information about the Settings dialog box, see the *SAS Data Loader for Hadoop: vApp Deployment Guide*. The dialog box contains certain fields for which the Hadoop administrator must provide values to the vApp user. These fields are as follows:

Table 5.1 *Settings Fields*

Field	Value
Host name	The host name that you create for Kerberos security. See "Client Configuration" on page 27.
User ID	The normal logon ID for the user.
Kerberos Realm	The name of the Kerberos realm or AD domain against which the user authenticates.
Kerberos configuration file	The location of the Kerberos configuration file.
Host keytab file	The location of the keytab generated for the host SPN. See "vApp" on page 28.
SAS server keytab file	The location of the keytab generated for the SAS server SPN. See "vApp" on page 28.
HTTP keytab file	The location of the keytab generated for the HTTP SPN. See "vApp" on page 28.

You must provide the Kerberos configuration and keytab files to the user.

Chapter 6

Updating the SAS Quality Knowledge Base (QKB)

SAS QKB Updates and Customization

Overview

SAS provides regular updates to the QKB. It is recommended that you update your QKB each time that a new one is released. For a listing of the latest enhancements to the QKB, see the What's New document on the SAS Quality Knowledge Base product documentation page at support.sas.com. To find this page, either search on the name SAS Quality Knowledge Base or locate the name in the product index and click the **Documentation** tab. Check the What's New for each QKB to determine which definitions have been added, modified, or deprecated, and to learn about new locales that might be supported. Contact your SAS software representative to order updated QKBs and locales. After obtaining the new QKB, copy it to the Hadoop NameNode (See "Copying the QKB to the Hadoop NameNode" on page 34) and use the same steps that you would to deploy a standard QKB.

The definitions delivered in the QKB are sufficient for performing most data quality operations. However, if you have DataFlux Data Management Studio, you can use the Customize feature to modify your QKB to meet specific needs. See your SAS representative for information about licensing DataFlux Data Management Studio.

If you want to customize your QKB, it is recommended that you customize your QKB on a local workstation, and then copy the customized QKB to the Hadoop NameNode for deployment. When updates to the QKB are required, merge your customizations into an updated QKB locally, and copy the updated, customized QKB to the Hadoop NameNode (See "Copying the QKB to the Hadoop NameNode" on page 34) for deployment. This enables you to deploy a customized QKB to the Hadoop cluster using the same steps that you would to deploy a standard QKB. Copying your customized QKB from a local workstation also means that you have a backup of the QKB on your local workstation. See the online Help provided with your SAS Quality Knowledge Base for information about how to merge any customizations that you have made into an updated QKB.

Updating

Kerberos Security Requirements

A Kerberos ticket (TGT) is required to deploy the QKB in a Kerberos environment.

To create the ticket, follow these steps:

1. Log on as root.
2. Change to the HDFS user.
3. Run kinit.
4. Exit to root.

The following is an example of commands used to obtain the ticket.

```
su - root
su - hdfs
kinit -kt hdfs.keytab hdfs
exit
```

Copying the QKB to the Hadoop NameNode

After you have obtained a QKB, you must copy it to the Hadoop NameNode. Copy the QKB to a temporary staging area, such as **/tmp/qkbstage**.

You can copy the QKB to the Hadoop NameNode by using a file transfer command like FTP or SCP, or by mounting the file system where the QKB is located on the Hadoop NameNode. You must copy the complete QKB directory structure.

SAS installation tools typically create a QKB in the following locations, where *qkb_product* is the QKB product name and *qkb_version* is the QKB version number:

- Windows 7: **C:\ProgramData\SAS\QKB*qkb_product**qkb_version***.

  For example:

  ```
  C:\ProgramData\SAS\QKB\CI\26
  ```

  Note: ProgramData is a hidden location.

- UNIX and Linux:**/opt/sas/qkb/*qkb_product*/*qkb_version***.

  For example:

  ```
  /opt/sas/qkb/ci/26
  ```

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 the command:

```
scp -r /opt/sas/qkb/ci/26 hmaster456:/tmp/qkbstage
```

Overview of the QKB_PUSH.SH Script

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 location.
- Write the installation or removal output to a log file.

The qkb_push.sh file is created in the ***EPInstallDir*/SASEPHome/bin** directory. You must execute qkb_push.sh from this directory. You can also use qkb_push.sh to deploy updated versions of the QKB. See "Copying the QKB to the Hadoop NameNode" on page 34

You suppress index creation or perform only index creation by using the **-i** and **-x** arguments. If users have a problem viewing QKB definitions from within SAS Data Loader, you might want to re-create the index file.

Note: Only one QKB and one index file are supported in the Hadoop framework at a time. For example, you cannot have a QKB for Contact Information and a QKB for Product Data in the Hadoop framework at the same 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. If a QKB or QKB index file already exists in the target location, the new QKB or QKB index file overwrites it.

Installing the QKB

Installing the QKB on the Hadoop cluster nodes performs the following two tasks:

- copies the specified QKB directory to a fixed location (**/opt/qkb/default**) on each of the Hadoop nodes.

  Note: Each Hadoop node requires approximately 8 GB of disk space for the QKB.

- generates an index file from the contents of the QKB and pushes this index file to HDFS. This index file, named default.idx, is created in the **/sas/qkb** directory in HDFS. The default.idx file provides a list of QKB definition and token names to SAS Data Loader.

To deploy the QKB to the cluster, run qkb_push.sh. You must run qkb_push.sh as the root user.

Run qkb_push.sh as follows:

```
cd EPInstallDir/SASEPHome/bin
./qkb_push.sh qkb_path
```

where *qkb_path* is the name of the directory on the NameNode to which you copied the QKB. For example, you might use the following:

```
./qkb_push.sh /tmp/qkbstage/version
```

The qkb_push.sh script automatically discovers all nodes of the cluster by default and deploys the QKB to those nodes. 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 "QKB_PUSH.SH Syntax" on page 36.

The qkb_push.sh script creates the following directories and files on each node on which it is executed:

```
EPInstallDir/opt/qkb/default/chopinfo
opt/qkb/default/dfx.meta
opt/qkb/default/grammar
opt/qkb/default/inst.meta
opt/qkb/default/locale
opt/qkb/default/phonetx
opt/qkb/default/regexlib
opt/qkb/default/scheme
opt/qkb/default/upgrade.40
opt/qkb/default/vocab
```

Verify that these directories and files have been copied to the nodes.

Check that the default.idx file was created in HDFS or MAPR by issuing the command:

```
hadoop fs -ls /sas/qkb
```

Removing the QKB

The QKB can be removed from the Hadoop cluster by executing the qkb_push.sh executable file with the **-r** argument. You must have root access to execute qkb_push.sh.

Note: If you are removing the entire in-database deployment, you must remove the QKB first.

Run qkb_push.sh as follows:

```
cd EPInstallDir/SASEPHome/bin
./qkb_push.sh -r
```

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.

Note: 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 "QKB_PUSH.SH Syntax" on page 36.

QKB_PUSH.SH Syntax

qkb_push.sh <*arguments*> *qkb_path*

<-?>

<-l *logfile*>

<-f *hostfile*>

<-h *hostname*>

<-v >

<-s *user-id*>

<-i >

<-x >

<-r >

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.

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" on page 37
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.
Requirement	If you specify more than one host, the host names must be separated by spaces.
Interaction	Use the -h argument in conjunction with the -r argument to remove the QKB from specific nodes.
Note	The -f and -h arguments are mutually exclusive.
Tip	Use the -host argument when new nodes are added to the cluster
See	"-r" on page 37
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 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.

-x
suppresses QKB index creation.

-r
removes the QKB from the Hadoop nodes and it removes the QKB index file from HDFS.

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 when the -h or -f argument is specified in conjunction with -r.

See "-f *hostfile*" on page 37

"-h *hostname hostname*" on page 37

Example

```
-r -h server1 server2 server3
-r -f /etc/hadoop/conf/slaves
-r -l logfile
```

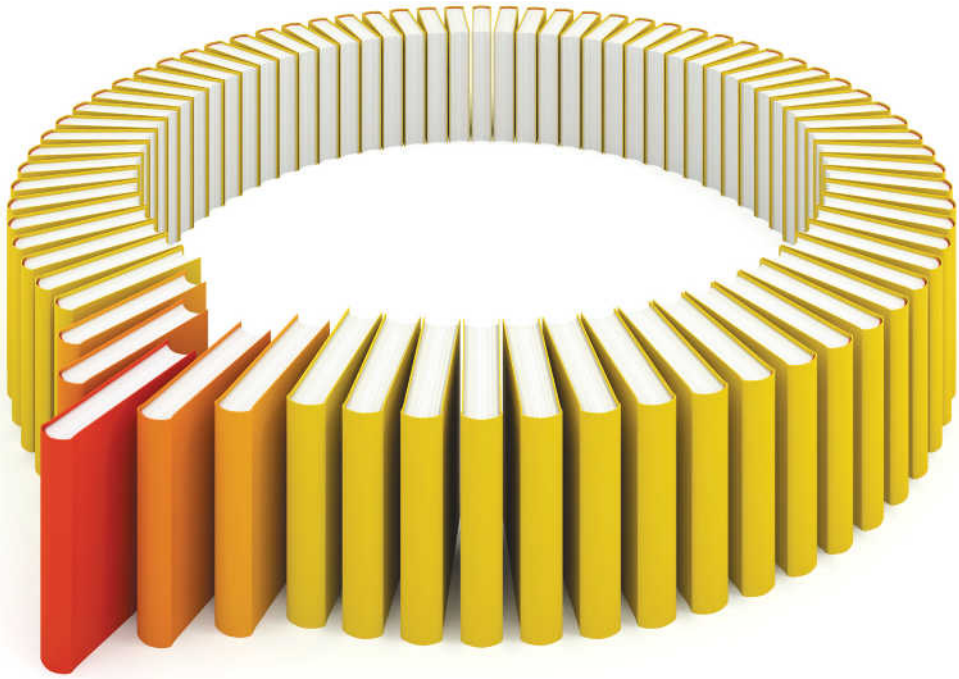
Recommended Reading

- *SAS Data Loader for Hadoop: User's Guide*
- *SAS Data Loader for Hadoop: vApp Deployment 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

Index



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.