SAS® High-Performance Analytics Infrastructure 2.1

Installation and Configuration Guide

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2013. *SAS® High-Performance Analytics Infrastructure 2.1: Installation and Configuration Guide*. Cary, NC: SAS Institute Inc.

SAS® High-Performance Analytics Infrastructure 2.1: Installation and Configuration Guide

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

Electronic book 1, July 2013

SAS® Publishing provides a complete selection of books and electronic products to help customers use SAS software to its fullest potential. For more information about our e-books, e-learning products, CDs, and hard-copy books, visit the SAS Publishing Web site at **support.sas.com/publishing** or call 1-800-727-3228.

Contents

What's New

What's New in Installation and Configuration for SAS High-Performance Analytics Infrastructure 2.1

Overview

The *SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide* explains how to install and initially configure the SAS High-Performance Analytics infrastructure. This infrastructure consists of the following products:

- SAS High-Performance Computing Management Console
- SAS High-Performance Deployment of Hadoop
- SAS High-Performance Analytics environment

This document contains new material on the following enhancements and changes to the SAS High-Performance Analytics infrastructure:

- support for Kerberos
- support for multiple data devices
- confining the scope of sudo

Support for Kerberos

During the installation of the SAS High-Performance Analytics environment, you can now enter a value in the mpirun prompt that enables support for Kerberos. For more information, see Table 5.1 on page 55.

Support for Multiple Data Devices

SAS provides a way to configure SAS High-Performance Deployment of Hadoop with multiple data devices. For more information, see “(Optional) Deploy with Multiple Data Devices” on page 45.

Confining the Scope of sudo

SAS supplies a security wrapper that you can use to limit the use of sudo commands to specific directories. The SAS High-Performance Deployment of Hadoop requires elevated privileges for temporary directories only. For more information, see “(Optional) Limit the Scope of sudo” on page 46.

Accessibility

Accessibility Features of the SAS Intelligence Platform Deployment Tools

Overview

For this release, the SAS 9.4 Intelligence Platform deployment tools have not been tested for compliance with U.S. Section 508 standards and W3C web content accessibility guidelines. If you have specific questions about the accessibility of SAS products, send them to **accessibility@sas.com** or call SAS Technical Support.

Recommended Reading

Here is the recommended reading list for this title:

- *Configuration Guide for SAS Foundation for Microsoft Windows for x64*, available at `http://support.sas.com/documentation/installcenter/en/ikfdtnwx6cg/66385/PDF/default/config.pdf`.
- *Configuration Guide for SAS Foundation for UNIX Environments*, available at `http://support.sas.com/documentation/installcenter/en/ikfdtnunxcg/66380/PDF/default/config.pdf`.
- *SAS Deployment Wizard and SAS Deployment Manager: User's Guide*, available at `http://support.sas.com/documentation/installcenter/en/ikdeploywizug/66034/PDF/default/user.pdf`.
- *SAS Guide to Software Updates*, available at `http://support.sas.com/documentation/cdl/en/whatsdiff/66129/PDF/default/whatsdiff.pdf`.
- *SAS High-Performance Computing Management Console: User's Guide*, available at `http://support.sas.com/documentation/solutions/hpainfrastructure/`.
- *SAS Intelligence Platform: Installation and Configuration Guide*, available at `http://support.sas.com/documentation/cdl/en/biig/63852/PDF/default/biig.pdf`.
- *SAS Intelligence Platform: Security Administration Guide*, available at `http://support.sas.com/documentation/cdl/en/bisecag/65011/PDF/default/bisecag.pdf`.

For a complete list of SAS books, go to support.sas.com/bookstore. If you have questions about which titles you need, please contact a SAS Book Sales Representative:

SAS Books
SAS Campus Drive
Cary, NC 27513-2414
Phone: 1-800-727-3228
Fax: 1-919-677-8166
E-mail: sasbook@sas.com
Web address: support.sas.com/bookstore

1

Introduction to Deploying the SAS High-Performance Analytics Infrastructure

What is the SAS High-Performance Analytics Infrastructure?

The SAS High-Performance Analytics infrastructure consists of tools for analytic tasks in a high-performance environment that is characterized by massively parallel processing (MPP) and symmetric multiprocessing (SMP) on a distributed database system.

The SAS High-Performance Analytics appliance provides a massively parallel computing environment supported by Message Passing Interface (MPI) combined with either a massively parallel distributed database management system (Teradata or EMC Greenplum) or a Hadoop Distributed File System on a x64-Linux platform.

The SAS High-Performance Analytics infrastructure consists of the following components:

- SAS High-Performance Analytics environment
- SAS High-Performance Deployment of Hadoop
- (Optional) SAS High-Performance Computing Management Console

Figure 1.1 *SAS High-Performance Analytics Infrastructure on a Supported Data Appliance*

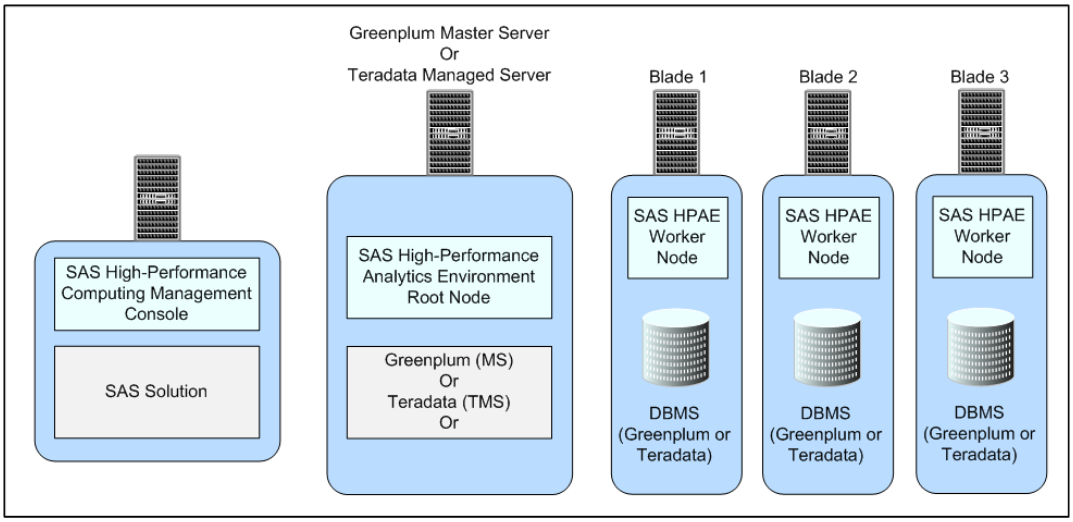
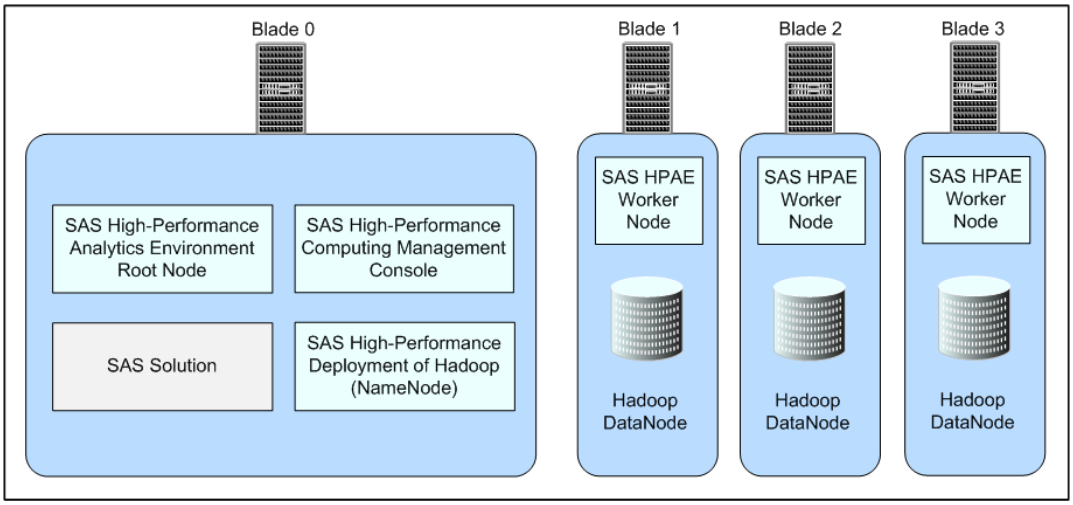


Figure 1.2 *SAS High-Performance Analytics Infrastructure on a Hadoop Machine Cluster*



The following products are among the various SAS solutions that use the SAS High-Performance Analytics infrastructure:

- SAS High-Performance Analytics Server
- SAS High-Performance Marketing Optimization
- SAS High-Performance Risk
- SAS Visual Analytics

What Is Covered in This Document?

This document covers tasks that are required after you and your SAS representative have decided what software you need and on what machines you will install the software. At this point, you can begin performing some pre-installation tasks, such as creating a SAS Software Depot if your site already does not have one and setting up the operating system user accounts that you will need.

By the end of this document, you will have deployed the SAS High-Performance Analytics environment, SAS High-Performance Computing Management Console, and SAS High-Performance Deployment of Hadoop (if your SAS solution relies on Hadoop).

You will then be ready to deploy your SAS solution (such as SAS Visual Analytics, SAS High-Performance Risk, and SAS High-Performance Analytics Server) on top of the SAS High-Performance Analytics infrastructure. For more information, see the documentation for your respective SAS solution.

Deploying the SAS High-Performance Analytics Infrastructure

Overview of Deploying the SAS High-Performance Analytics Infrastructure

The following list summarizes the steps required to install and configure the SAS High-Performance Analytics infrastructure:

1. Review additional documentation.
2. Prepare your system.
3. Create a SAS Software Depot.
4. (Optional) Deploy SAS High-Performance Computing Management Console.
5. (Optional) Install and configure SAS High-Performance Deployment of Hadoop.
6. Deploy the SAS High-Performance Analytics environment.
7. Configure your data storage.

The following sections provide a brief description of each of these tasks. Subsequent chapters in the guide provide the step-by-step instructions that you will need to perform them.

Step 1: Review Additional Documentation

It is very important to review all the different documents associated with deploying your SAS software. There can be late-breaking information, or instructions specific to a particular configuration might be too narrow for inclusion in this document.

Your review should include these documents:

- QuickStart Guide

  This document is shipped with your SAS software. Follow its instructions.

  The QuickStart Guides are also available at **http://support.sas.com/documentation/installcenter/94/unx/index.html**

- software order e-mail (SOE)

  This e-mail is sent to your site to announce the software and detail the order. It also enumerates the initial installation steps and, for SAS 9.3, contains instructions for using Electronic Software Delivery (ESD), if applicable. The SID file also contains your site's SAS license (SETINIT).

- SAS order information (SOI)

  After you download your order to an existing SAS Software Depot, you can use the SAS order information (SOI) file to determine what products were in your order and when the order was placed. The SOI is in your SAS Software Depot in **install_doc/***order-number***/soi.html**.

- SAS Software Summary

  In the same depot location as the SOI, the SAS software summary is a more detailed list of the software that is included in your order. Unlike the SAS order information sheet, which lists only the software that you have specifically ordered, this document also describes the included software that supports your order. The software summary is in your SAS Software Depot in **install_doc/***order-number***/ordersummary.html**.

- system requirements

  Refer to the system requirements for your SAS solution, available at **http://support.sas.com/resources/sysreq/index.html**.

- SAS Notes

  Outstanding SAS Notes for alert status installation problems are available at

  http://support.sas.com/notes/index.html.

Step 2: Prepare Your System

Preparing your system includes tasks such as creating a list of machine names in your grid hosts file. Setting up passwordless SSH is required, as well as considering system umask settings. You must determine which operating system users you will require to install, configure, and run the SAS High-Performance Analytics infrastructure. Also, you will need to designate ports for the various SAS components that you are deploying.

For more information, see Chapter 2, "Preparing Your System to Deploy the SAS High-Performance Analytics Infrastructure," on page 9.

Step 3: Create a SAS Software Depot

Create a SAS Software Depot, which is a special file system used to deploy your SAS software. The depot contains the SAS Deployment Wizard—the program used to install and initially configure most SAS software—one or more deployment plans, a SAS installation data file, order data, and product data.

Note: If you have elected to receive SAS through Electronic Software Delivery, a SAS Software Depot is automatically created for you.

For more information, see "Creating a SAS Software Depot" in the *SAS Intelligence Platform: Installation and Configuration Guide*, available at `http://support.sas.com/documentation/cdl/en/biig/63852/HTML/default/p03intellplatform00installgd.htm`.

Step 4: (Optional) Deploy SAS High-Performance Computing Management Console

SAS High-Performance Computing Management Console is an optional web application tool that eases the administrative burden on multiple machines in a distributed computing environment.

When you are creating operating system accounts and passwordless SSH on all machines in the cluster or on blades across the appliance, the management console enables you to perform these tasks from one location.

You can also manage CPU and memory resources across the cluster through management console support for CGroups that is built in to Linux.

For more information, see Chapter 3, “Deploying SAS High-Performance Computing Management Console,” on page 23.

Step 5: (Optional) Install and Configure SAS High-Performance Deployment of Hadoop

If your site is using Hadoop, then you will install and configure SAS High-Performance Deployment of Hadoop, which consists of a NameNode and DataNodes. The product is installed by a self-extracting shell script.

For more information, see Chapter 4, “Installing and Configuring SAS High-Performance Deployment of Hadoop,” on page 39.

Step 6: Deploy the SAS High-Performance Analytics Environment

The SAS High-Performance Analytics environment consists of a root node and worker nodes. The product is installed by a self-extracting shell script.

The root node is deployed on the grid host. A worker node is installed on each remaining machine in the cluster or database appliance.

For more information, see Chapter 5, “Deploying the SAS High-Performance Analytics Environment,” on page 51.

Step 7: Configure Your Data Storage

Depending on which data provider you plan to use with SAS, there are certain configuration tasks that you will need to complete on the Hadoop machine cluster or the Greenplum or Teradata appliance.

For more information, see Chapter 6, “Configuring Your Data Storage,” on page 59.

2

Preparing Your System to Deploy the SAS High-Performance Analytics Infrastructure

SAS High-Performance Analytics Infrastructure Deployment Process Overview

Setting up operating system users is the second of seven steps required to install and configure the SAS High-Performance Analytics infrastructure.

1. Review additional documentation.
2. **2. Prepare your system.**
3. Create a SAS Software Depot.
4. (Optional) Deploy SAS High-Performance Computing Management Console.
5. (Optional) Install and configure SAS High-Performance Deployment of Hadoop.
6. Deploy the SAS High-Performance Analytics environment.
7. Configure your data storage.

System Settings for the SAS High-Performance Analytics Infrastructure

Understand the system requirements for a successful SAS High-Performance Analytics infrastructure deployment before you begin. The lists that follow offer recommended settings for the analytics infrastructure on every machine in the cluster or blade in the data appliance:

- Modify **/etc/ssh/sshd_config** with the following setting:

  ```
  MaxStartups 1000
  ```

- Modify **/etc/security/limits.conf** with the following settings:
  - `soft nproc 65536`
  - `hard nproc 65536`
  - `soft nofile 350000`
  - `hard nofile 350000`

- Modify **/etc/security/limits.d/90-nproc.conf** with the following setting:

  ```
  soft nproc 65536
  ```

- Modify **/etc/sysconfig/cpuspeed** with the following setting:

  ```
  GOVERNOR=performance
  ```

- The SAS High-Performance Analytics components require approximately 580 MB of disk space. SAS High-Performance Deployment of Hadoop requires approximately 300 MB of disk space for the software. This estimate does not include the disk space that is needed for storing data that is added to Hadoop Distributed File System (HDFS) for use by the SAS High-Performance Analytics environment.

  For more information, refer to the system requirements for your SAS solution, available at **http://support.sas.com/resources/sysreq/index.html**.

List the Machines in the Cluster or Appliance

Before the SAS High-Performance Analytics infrastructure can be installed on the machines in the cluster, you must create a file that lists all of the host names of the machines in the cluster.

On blade 0, known as the Master Server (Greenplum) or the Managed Server (Teradata), create an **`/etc/gridhosts`** file for use by SAS High-Performance Computing Management Console, SAS High-Performance Deployment of Hadoop, and the SAS High-Performance Analytic environment. (The grid hosts file is copied to the other machines in the cluster during the installation process.) If additional machines are used outside of the cluster for the SAS solution server, the SAS middle tier, or SAS High-Performance Computing Management Console, then these machines must each contain a copy of **`/etc/gridhosts`**. For more information, see "Deploying SAS High-Performance Computing Management Console" on page 24 before you start the installation.

You can use short names or fully qualified domain names so long as the host names in the file resolve to IP addresses. The long and short host names for each node must be resolvable from each node in the environment. The host names listed in the file must be in the same DNS domain and sub-domain. These host names are used for Message Passing Interface (MPI) communication and SAS High-Performance Deployment of Hadoop network communication.

The *root node* is listed first. This is also the machine that is configured as the following, depending on your data provider:

- SAS High-Performance Deployment of Hadoop: NameNode (blade 0)
- Greenplum Data Computing Appliance: Master Server
- Teradata: Managed Server

The following lines are an example of the file contents:

```
grid001
```

```
grid002
grid003
grid004
...
```

TIP You can use SAS High-Performance Computing Management Console to create and manage your grid hosts file. For more information, see *SAS High-Performance Computing Management Console: User's Guide* available at `http://support.sas.com/documentation/onlinedoc/va/index.html`.

Review Passwordless Secure Shell Requirements

Passwordless Secure Shell (SSH) is required on all machines in the cluster or on the data appliance for the following user accounts:

- root user account

  The root account must run SAS High-Performance Computing Management Console and the simultaneous commands (for example, **`simsh`**, and **`simcp`**). For more information about management console user accounts, see “Preparing to Install SAS High-Performance Computing Management Console” on page 14.

- Hadoop user account

  For more information about Hadoop user accounts, see “Preparing to Install and Configure SAS High-Performance Deployment of Hadoop” on page 15.

- SAS High-Performance Analytics environment user account

  For more information about the environment’s user accounts, see “Preparing to Deploy the SAS High-Performance Analytics Environment” on page 19.

Preparing to Install SAS High-Performance Computing Management Console

User Account Considerations for SAS High-Performance Computing Management Console

SAS High-Performance Computing Management Console is installed from an RPM package and must be installed and configured with the root user ID. The root user account must have passwordless secure shell (SSH) access between all the machines in the cluster. The console includes a web server. The web server is started with the root user ID, and it runs as the root user ID.

The reason that the web server for the console must run as the root user ID is that the console can be used to add, modify, and delete operating system user accounts from the local passwords database (`/etc/passwd` and `/etc/shadow`). Only the root user ID has Read and Write access to these files.

Be aware that you do not need to log on to the console with the root user ID. In fact, the console is typically configured to use console user accounts. Administrators can log on to the console with a console user account that is managed by the console itself and does not have any representation in the local passwords database or whatever security provider the operating system is configured to use.

Management Console Requirements

Before you install SAS High-Performance Computing Management Console, make sure that you have performed the following tasks:

- Make sure that the Korn shell is installed.
- Make sure that the Perl extension perl-Net-SSLeay is installed.
- For PAM authentication, make sure that the Authen::PAM PERL module is installed.

- Create the list of all the cluster machines in the **/etc/gridhosts** file. You can use short names or fully qualified domain names so long as the host names in the file resolve to IP addresses. These host names are used for Message Passing Interface (MPI) communication and Hadoop network communication. For more information, see "List the Machines in the Cluster or Appliance" on page 12.
- Locate the software.

  Make sure that your SAS Software Depot has been created. (For more information, see "Creating a SAS Software Depot" in the *SAS Intelligence Platform: Installation and Configuration Guide*, available at **http://support.sas.com/documentation/cdl/en/biig/63852/HTML/default/p03intellplatform00installgd.htm**.)

Preparing to Install and Configure SAS High-Performance Deployment of Hadoop

User Accounts for SAS High-Performance Deployment of Hadoop

The account with which you deploy Hadoop must have passwordless secure shell (SSH) access between all the machines in the cluster.

TIP Although the Hadoop installation program can run as any user, you might find it easier to run **hadoopInstall** as root so that it can set permissions and ownership of the Hadoop data directories for the user account that runs Hadoop.

An operating system user ID is required to run the Hadoop applications on the machines in the cluster. This user ID must exist on all the machines in the cluster and must be configured for passwordless SSH.

The SAS High-Performance Deployment of Hadoop installation program checks to see whether the user account and group that you specify is present. If this user account and group is not present, then the program creates the user account and group on each machine in the cluster before installing SAS High-Performance Deployment of Hadoop.

If you do not already have an account that meets the requirements, you can use SAS High-Performance Computing Management Console to add the appropriate user ID.

As a convention, this document uses an account and group named **hadoop** when describing how to deploy and run SAS High-Performance Deployment of Hadoop.

Note: To properly run Hadoop, you must make certain modifications to the **/etc/sudoers** file on each machine in the cluster. For more information, see "Modify the sudoers File" on page 16.

If your site has a requirement for a reserved UID and GID for the Hadoop user account, then create the user and group on each machine before continuing with the installation.

Note: We recommend that you install SAS High-Performance Computing Management Console before setting up the user accounts that you will need for the rest of the SAS High-Performance Analytics infrastructure. The console enables you to easily manage user accounts across the machines of a cluster. For more information, see "Create the First User Account and Propagate the SSH Key" on page 34.

SAS High-Performance Deployment of Hadoop is installed from a TAR.GZ file. An installation and configuration program, **hadoopInstall**, is available after the archive is extracted.

SAS High-Performance Deployment of Hadoop includes a security feature that sets file system ownership and permissions for the files that are used as blocks in the Hadoop Distributed File System (HDFS). The files are subject to the owner, group, and other mode permissions that are commonly understood for POSIX file systems. The files are also subject to the umask setting for the user that writes the blocks. Because the owner and group (and possibly the mode) are changed on the files, the user ID that is used to run the Hadoop server process must have Read (and Write) permission to the files.

The permission mode on files is set either through the user's umask setting or as a data set option when users use the SAS Data in HDFS engine to distribute data in HDFS.

Modify the sudoers File

In order to format the Hadoop NameNode (discussed later in this document) and to properly run Hadoop, you must make certain modifications to the **/etc/sudoers** file on each machine in the cluster.

How you modify the sudoers file is dependent on whether you are implementing the security wrapper. (For more information, see "(Optional) Limit the Scope of sudo" on page 46.)

- If you are using the security wrapper, then add the absolute path to the wrapper script in the **/etc/sudoers** file on each machine in the cluster.

  For example:

  /hadoop/hadoop/my_security_wrapper.pl

- If you are *not* using the security wrapper, do one of the following:

  - If you are running Hadoop as the *root* user, make sure that the following lines are present in the **/etc/sudoers** file on each machine in the cluster:

    ```
    Defaults:root !requiretty
    root ALL=NOPASSWD:/bin/mkdir,/bin/chown
    ```

  - If you are running Hadoop as a *non-root* user, make sure that the following lines are present in the **/etc/sudoers** file on each machine in the cluster:

    ```
    Defaults:non-root-user-ID !requiretty
    non-root-user-ID ALL=NOPASSWD:/bin/mkdir,/bin/chown
    ```

TIP You can issue a single **simcp** command to propagate one sudoers file across all machines in the cluster. The **simcp** and **simsh** commands are available with SAS High-Performance Computing Management Console. For more information, see Appendix 3, "SAS High-Performance Analytics Infrastructure Command Reference," on page 85.

Install a Java Runtime Environment

SAS High-Performance Deployment of Hadoop requires a Java Runtime Environment (JRE) or Java Development Kit (JDK) on every machine in the cluster. The path to the Java executable must be the same on all of the machines in the cluster. If this requirement is already met, make a note of the path and proceed to installing SAS High-Performance Deployment of Hadoop.

If the requirement is not met, then install a JRE or JDK on the machine that is used as the grid host. If you have installed SAS High-Performance Computing Management Console, then you can use the **simsh** and **simcp** commands to copy the files to the other machines in the cluster.

Example Code 2.1 *Sample simsh and simcp Commands*

```
/opt/webmin/utilbin/simsh mkdir /opt/java
/opt/webmin/utilbin/simcp /opt/java/jdk1.6.0_31 /opt/java
```

For information about the supported Java version, see **http://wiki.apache.org/hadoop/HadoopJavaVersions**. SAS High-Performance Deployment of Hadoop uses the Apache Hadoop 0.23.1 version.

Plan for Hadoop Directories

The following table lists the default directories where the SAS High-Performance Deployment of Hadoop stores content:

Table 2.1 *Default SAS High-Performance Deployment of Hadoop Directory Locations*

Default Directory Location	Description
hadoop-name	The **hadoop-name** directory is the location on the file system where the NameNode stores the namespace and transactions logs persistently. This location is formatted by Hadoop during the configuration stage.
hadoop-data	The **hadoop-data** directory is the location on the file system where the DataNodes store data in blocks.
hadoop-local	The **hadoop-local** directory is the location on the file system where temporary MapReduce data is written. MapReduce is not used by the SAS High-Performance Analytics environment, but specifying a location is a requirement of Hadoop.
hadoop-system	The **hadoop-system** directory is the location on the file system where the MapReduce framework writes system files. MapReduce is not used by the SAS High-Performance Analytics environment, but specifying a location is a requirement of Hadoop.

Note: These Hadoop directories must reside on local storage. The exception is the **`hadoop-data`** directory, which can be on a storage area network (SAN). Network attached storage (NAS) devices are not supported.

You create the Hadoop installation directory on the NameNode machine. The installation script prompts you for this Hadoop installation directory and the names for each of the subdirectories (listed in Table 2.1) which it creates for you on every machine in the cluster.

Especially in the case of the data directory, it is important to designate a location that is large enough to contain all of your data. If you want to use more than one data device, see "(Optional) Deploy with Multiple Data Devices" on page 45.

Preparing to Deploy the SAS High-Performance Analytics Environment

User Accounts for the SAS High-Performance Analytics Environment

This topic describes the user account requirements for deploying and running the SAS High-Performance Analytics environment:

- Installation and configuration must be run with the same user account.
- The installer account must have passwordless secure shell (SSH) access between all the machines in the cluster.

Note: We recommend that you install SAS High-Performance Computing Management Console before setting up the user accounts that you will need for the rest of the SAS High-Performance Analytics infrastructure. The console enables you to easily manage user accounts across the machines of a cluster. For more information, see "User Account Considerations for SAS High-Performance Computing Management Console" on page 14.

The SAS High-Performance Analytics environment uses a shell script installer. You can use a SAS installer account to install this software if the user account meets the following requirements:

- The SAS installer account has Write access to the directory that you want to use and Write permission to the same directory path on every machine in the cluster.
- The SAS installer account is configured for passwordless SSH on all the machines in the cluster.

The root user ID can be used to install the SAS High-Performance Analytics environment, but it is not a requirement. When users start a process on the machines in the cluster with SAS software, the process runs under the user ID that starts the process.

Consider Umask Settings

The SAS High-Performance Analytics environment installation script (described in a later section) prompts you for a umask setting. Its default is no setting.

If you do not enter any umask setting, then jobs, servers, and so on, that use the analytics environment create files with the user's pre-existing umask set on the operating system. If you set a value for umask, then that umask is used and overrides each user's system umask setting.

Entering a value of 027 ensures that only users in the same operating system group can read these files.

Note: Remember that the account used to run the LASRMonitor process (by default, sas) must be able to read the table and server files in `/opt/VADP/var` and any other related subdirectories.

For more information about using umask, refer to your Linux documentation.

Pre-installation Ports Checklist for SAS

While you are creating operating system user accounts and groups, you need to review the set of ports that SAS will use by default. If any of these ports is unavailable, select an alternate port, and record the new port on the ports pre-installation checklist that follows.

The following checklist indicates what ports are used for SAS by default and gives you a place to enter the port numbers that you will actually use.

We recommend that you document each SAS port that you reserve in the following standard location on each machine: `/etc/services`. This practice will help avoid port conflicts on the affected machines.

Note: These checklists are superseded by more complete and up-to-date checklists that can be found at `http://support.sas.com/installcenter/plans`. This website also contains a corresponding deployment plan and an architectural diagram. If you are a SAS solutions customer, consult the pre-installation checklist provided by your SAS representative for a complete list of ports that you must designate.

Table 2.2 *Pre-installation Checklist for SAS Ports*

SAS Component	Default Port	Data Direction	Actual Port
Hadoop Service on the NameNode	15452	Inbound	
Hadoop Service on the DataNode	15453	Inbound	
Hadoop DataNode Address	50010	Inbound	
Hadoop DataNode IPC Address	50020	Inbound	
SAS High-Performance Computing Management Console server	10020	Inbound	
Hadoop JobTracker	50030	Inbound	

SAS Component	Default Port	Data Direction	Actual Port
Hadoop TaskTracker	50060	Inbound	
Hadoop Name Node web interface	50070	Inbound	
Hadoop DataNode HTTP Address	50075	Inbound	
Hadoop Secondary NameNode	50090	Inbound	
Hadoop Name Node Backup Address	50100	Inbound	
Hadoop Name Node Backup HTTP Address	50105	Inbound	
Hadoop Name Node HTTPS Address	50470	Inbound	
Hadoop DataNode HTTPS Address	50475	Inbound	
SAS High-Performance Deployment of Hadoop	54310	Inbound	
SAS High-Performance Deployment of Hadoop	54311	Inbound	

Configuring SAS Enterprise Miner with SAS High-Performance Analytic Environment

To configure SAS Enterprise Miner to run with the SAS High-Performance Analytic environment, you must install HPTMINE language files on each machine in the cluster or each machine in the database appliance. For more information, see *the SAS Text Miner: High-Performance Procedures* available at `http://support.sas.com/documentation/onlinedoc/txtminer/index.html`.

3

Deploying SAS High-Performance Computing Management Console

SAS High-Performance Analytics Infrastructure Deployment Process Overview

Installing and configuring SAS High-Performance Computing Management Console is an optional fourth of seven steps required to install and configure the SAS High-Performance Analytics infrastructure.

1. Review additional documentation.
2. Prepare your system.
3. Create a SAS Software Depot.
4. **(Optional) Deploy SAS High-Performance Computing Management Console.**
5. (Optional) Install and configure SAS High-Performance Deployment of Hadoop.
6. Deploy the SAS High-Performance Analytics environment.
7. Configure your data storage.

Benefits of SAS High-Performance Computing Management Console

Passwordless SSH is required to start and stop SAS LASR Analytic Servers and to load tables. For some SAS solutions, such as SAS High-Performance Risk and SAS High-Performance Analytic Server, passwordless SSH is required to run jobs on the machines in the cluster.

Also, users of some SAS solutions must have an operating system (external) account on all the machines in the cluster and must have the key distributed across the cluster. For more information, see “Create the First User Account and Propagate the SSH Key” on page 34.

SAS High-Performance Computing Management Console enables you to perform these tasks from one location. When you create *new* user accounts using SAS High-Performance Computing Management Console, the console propagates the public key across all the machines in the cluster in a single operation.

Finally, SAS High-Performance Computing Management Console enables you to easily manage distributed CPU and memory resources. The management console relies on support for the CGroups feature that is provided by the Linux kernel and CGroups libraries. For more information, see *SAS High-Performance Computing Management Console: User's Guide*, available at `http://support.sas.com/documentation/solutions/hpainfrastructure/`.

Overview of Deploying SAS High-Performance Computing Management Console

Deploying SAS High-Performance Computing Management Console requires installing and configuring components on a machine other than the Greenplum or Teradata appliance. In this document, the management console is deployed on the machine where the SAS Solution is deployed.

Figure 3.1 *Management Console Deployed with a Data Appliance*

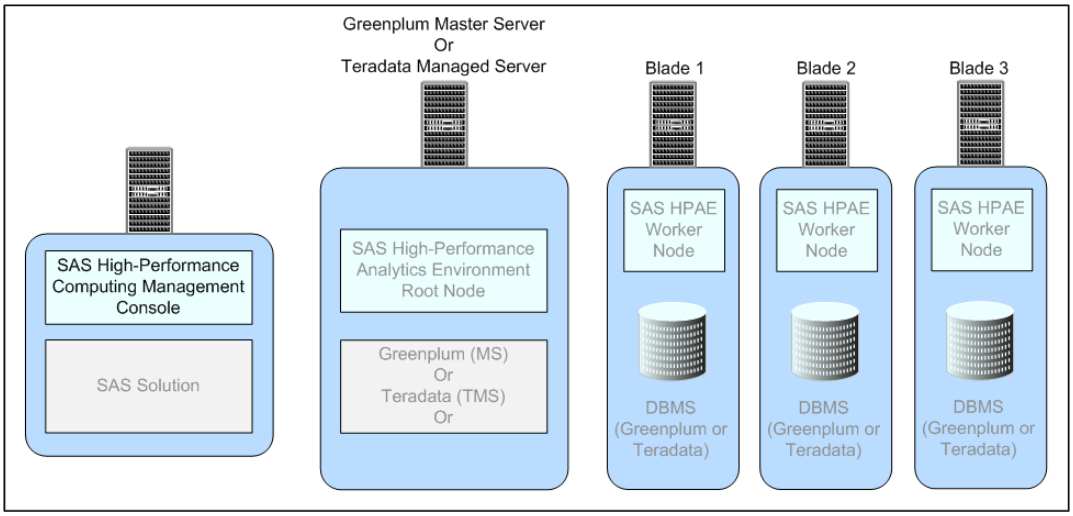
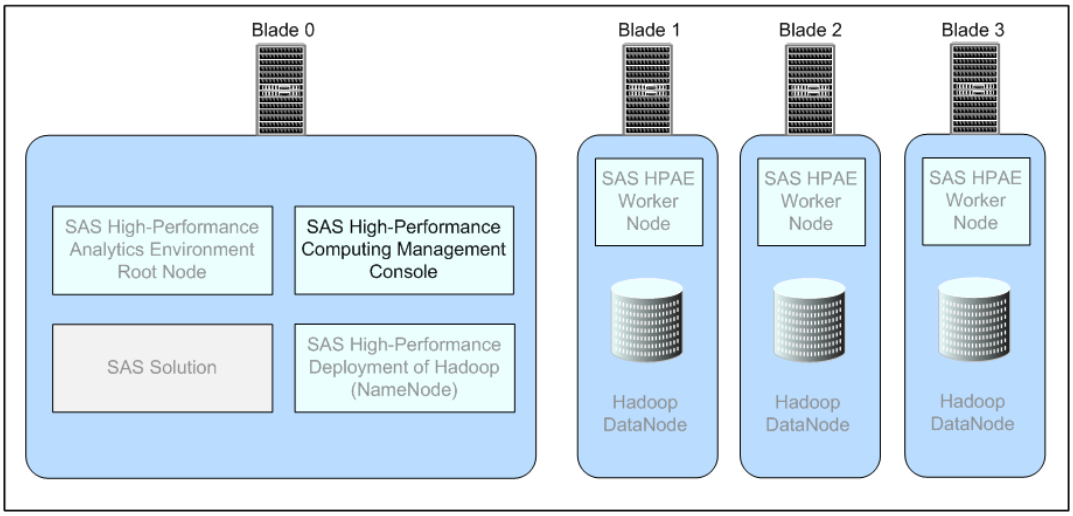


Figure 3.2 *Management Console Deployed on a Hadoop Machine Cluster*



Install SAS High-Performance Computing Management Console

To install and initially configure SAS High-Performance Computing Management Console, follow these steps:

Note: For information about updating the console, see "Updating the SAS High-Performance Analytics Infrastructure" on page 81.

1 Make sure that you have reviewed all of the information contained in the section "Preparing to Install SAS High-Performance Computing Management Console" on page 14.

2 Log on to the target machine as root.

3 In your SAS Software Depot, locate the **`standalone_installs/SAS_High-Performance_Management_Console/2_1/Linux_for_x64`** directory.

4 Enter the following command:

```
rpm -ivh sashpcmc*
```

5 Proceed to the topic "Configure SAS High-Performance Computing Management Console" on page 27.

Configure SAS High-Performance Computing Management Console

After installing SAS High-Performance Computing Management Console, you must configure it. This is done with the setup script.

1 Log on to the SAS Visual Analytics server and middle tier machine (blade 0) as root.

2 Run the setup script by entering the following command:

```
/opt/webmin/utilbin/setup
```

Answer the prompts that follow.

```
Enter the username for initial login to SAS HPC MC below.
This user will have rights to everything in the SAS HPC MC and
can either be an OS account or new console user. If an OS account
exists for the user, then system authentication will be used. If
an OS account does not exist, you will be prompted for a password.
```

3 Enter the user name for the initial login.

```
Creating sas using system authentication
Use SSL\HTTPS (yes|no)
```

4 If you want to use Secure Sockets Layer (SSL) when running the console, enter **yes**. Otherwise, enter **no**.

5 If you chose not to use SSL, then skip to Step 7 on page 28. Otherwise, the script prompts you to use a pre-existing certificate and key file or to create a new one.

```
Use existing combined certificate and key file or create a new one (file|create)?
```

6 Make one of two choices:

- Enter **create** to have the script generate the combined private key and SSL certificate file for you.

  The script displays output of the **openssl** command that it uses to create the private key pair for you.

- Enter **file** to supply the path to a valid private key pair.

  When prompted, enter the absolute path for the combined certificate and key file.

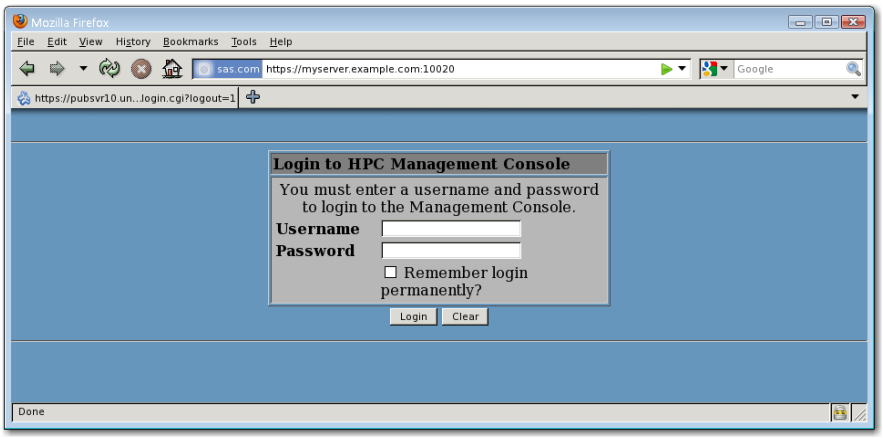
7 To start the SAS High-Performance Computing Management Console server, enter the following command from any directory:

```
service sashpcmc start
```

8 Open a web browser and, in the address field, enter the fully qualified domain name for the blade 0 host followed by port 10020.

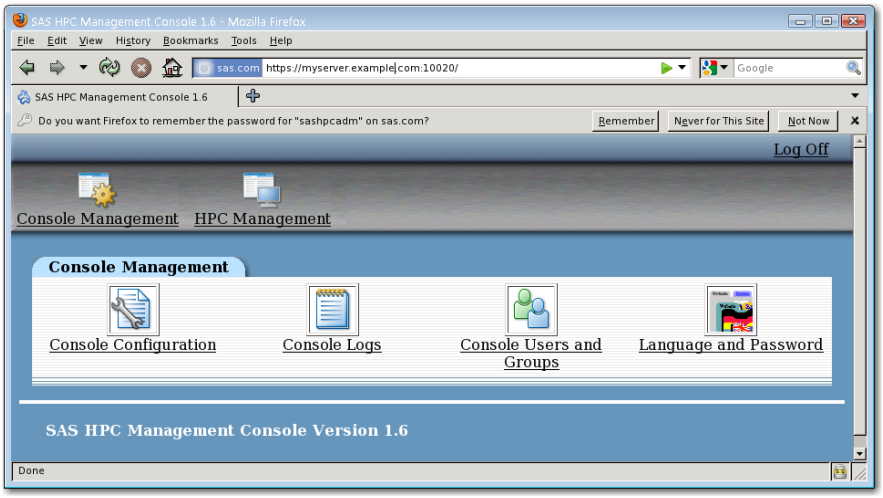
For example: `https://myserver.example.com:10020`

The Login page appears.



9 Log on to SAS High-Performance Computing Management Console using the credentials that you specified in Step 2.

The Console Management page appears.



Create the Installer Account and Propagate the SSH Key

The user account needed to start and stop server instances and to load and unload tables to those servers must be configured with passwordless secure shell (SSH).

To reduce the number of operating system (external) accounts, it can be convenient to use the SAS Installer account for both of these purposes.

Implementing passwordless SSH requires that the public key be added to the authorized_keys file across all machines in the cluster. When you create user accounts using SAS High-Performance Computing Management Console, the console propagates the public key across all the machines in the cluster in a single operation.

To create an operating system account and propagate the public key, follow these steps:

1 Make sure that the SAS High-Performance Computing Management Console server is running. While logged on as the root user, enter the following command from any directory:

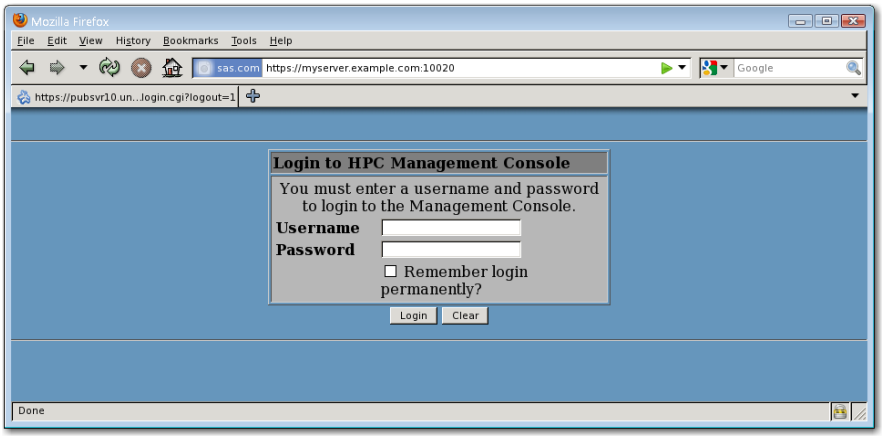
```
service sashpcmc status
```

(If you are logged on as a user other than the root user, the script returns the message `sashpcmc is stopped`.) For more information, see To start the SAS High-Performance Computing Management Console server on page 28.

2 Open a web browser and, in the address field, enter the fully qualified domain name for the blade 0 host followed by port 10020.

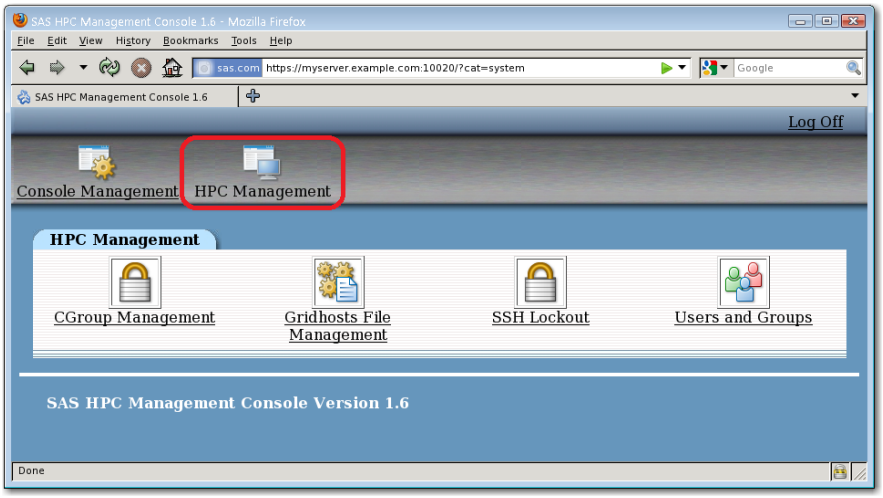
For example: **`http://myserver.example.com:10020`**

The Login page appears.



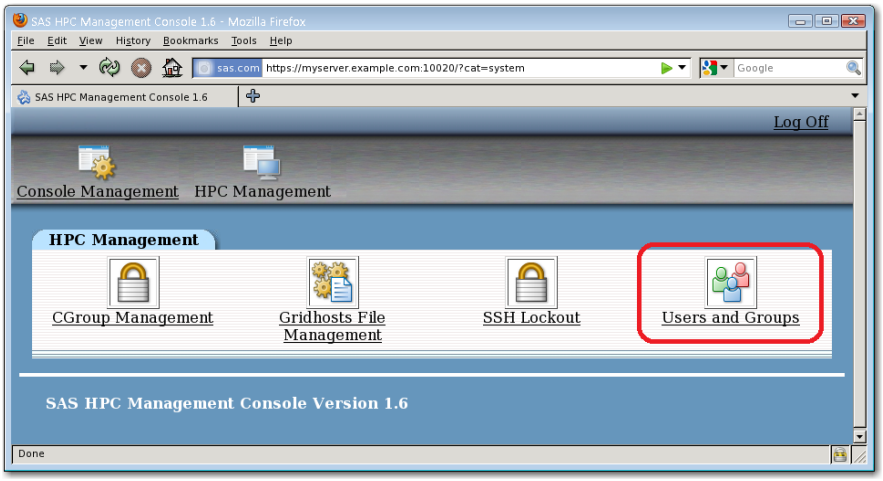
3 Log on to SAS High-Performance Computing Management Console.

The Console Management page appears.



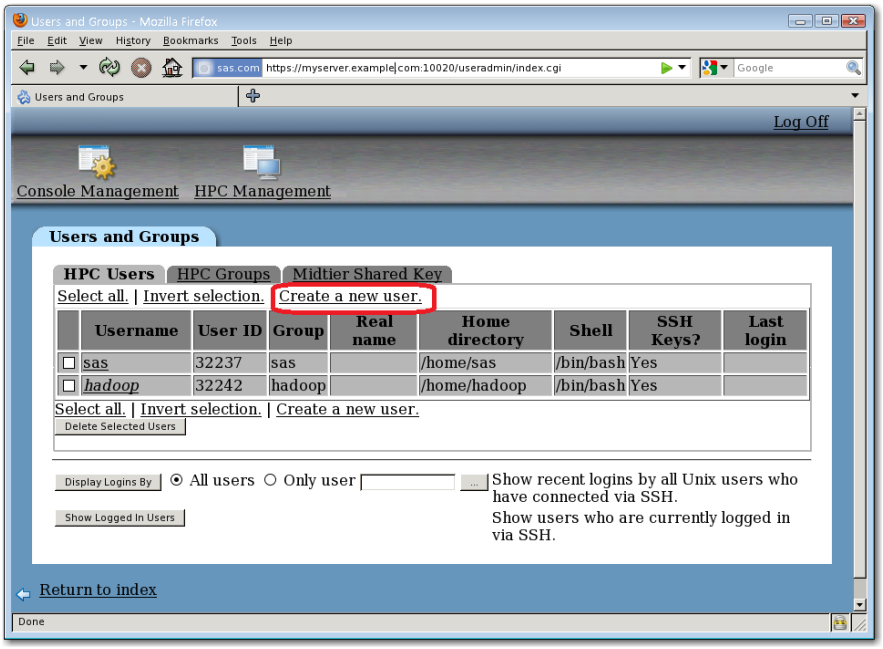
4 Click **HPC Management**.

The HPC Management page appears.



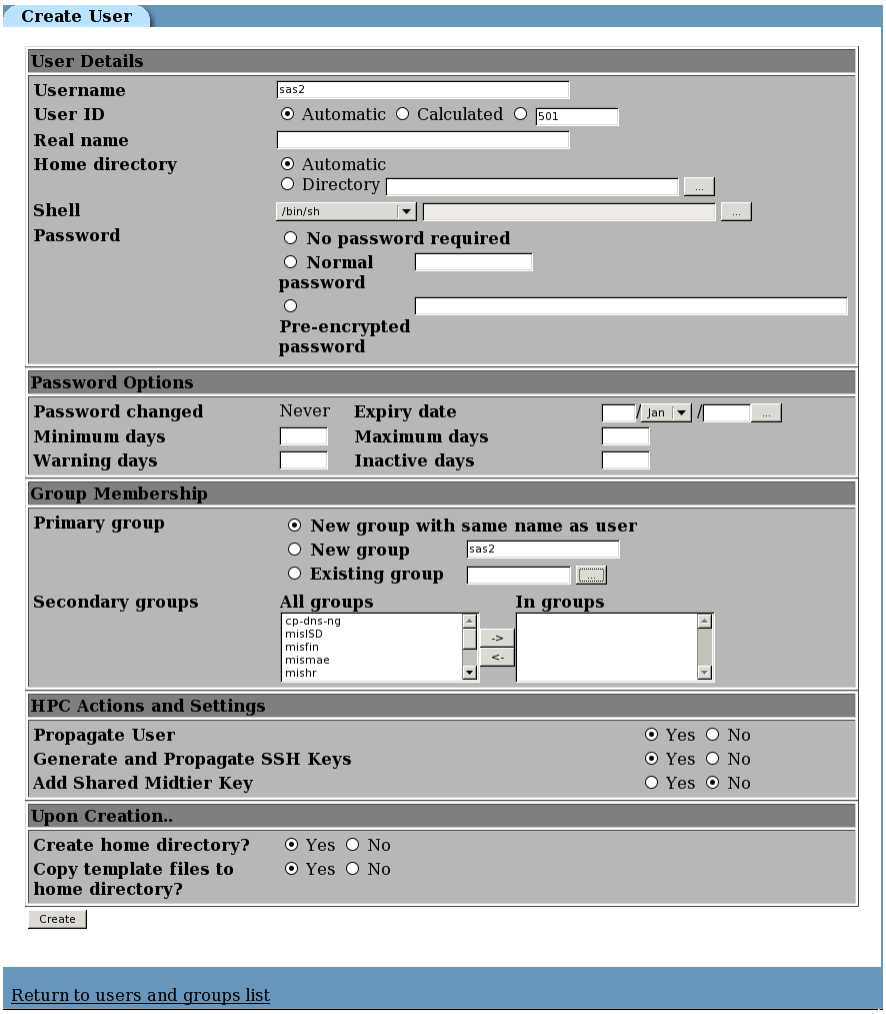
5 Click **Users and Groups**.

The Users and Groups page appears.



6 Click **Create a new user**.

The Create User page appears.



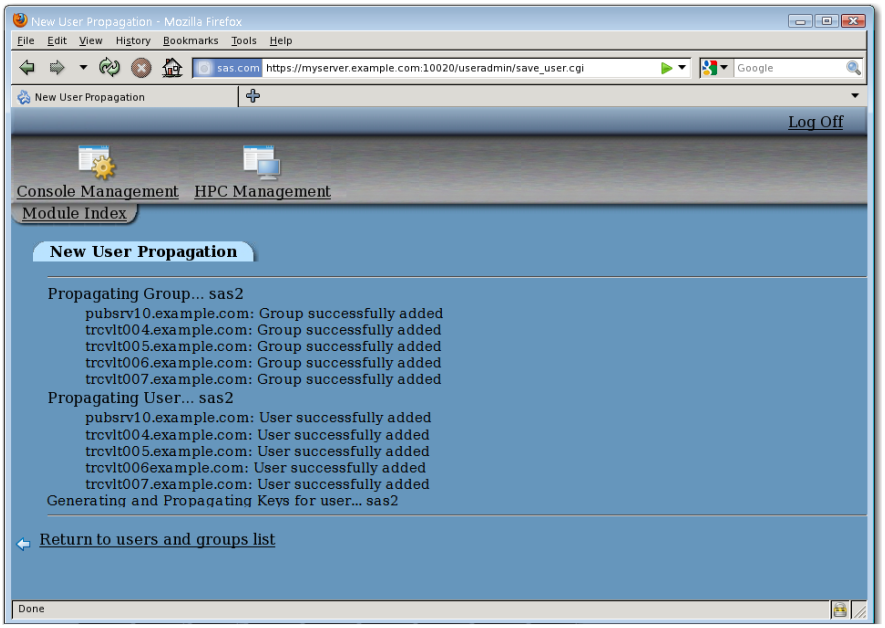
7 Enter information for the new user, using the security policies in place at your site.

Be sure to choose **Yes** for the following:

- **Propagate User**
- **Generate and Propagate SSH Keys**

When you are finished making your selections, click **Create**.

The New User Propagation page appears and lists the status of the create user command. Your task is successful if you see output similar to the following figure.



Create the First User Account and Propagate the SSH Key

Depending on their configuration, some SAS solution users must have an operating system (external) account on all the machines in the cluster. Furthermore, the public key might be distributed on each cluster machine in order for their secure shell (SSH) access to operate properly. SAS High-Performance Computing Management Console enables you to perform these two tasks from one location.

To create an operating system account and propagate the public key for SSH, follow these steps:

1 Make sure that the SAS High-Performance Computing Management Console server is running. Enter the following command from any directory:

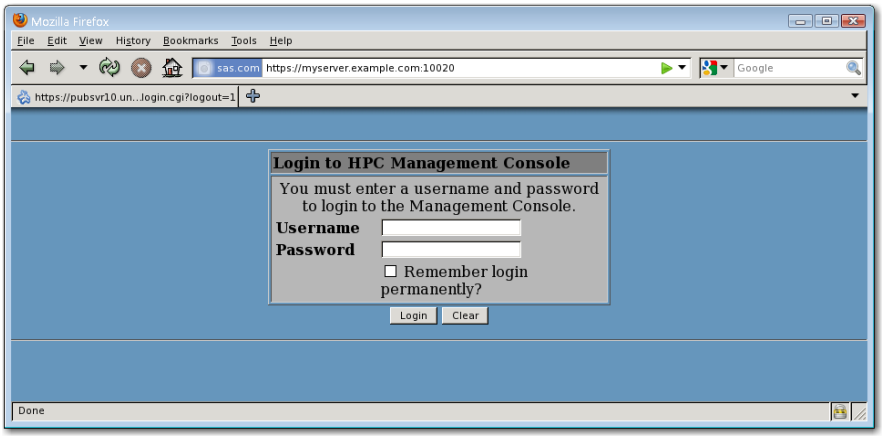
```
service sashpcmc status
```

For more information, see To start the SAS High-Performance Computing Management Console server on page 28.

2 Open a web browser and, in the address field, enter the fully qualified domain name for the blade 0 host followed by port 10020.

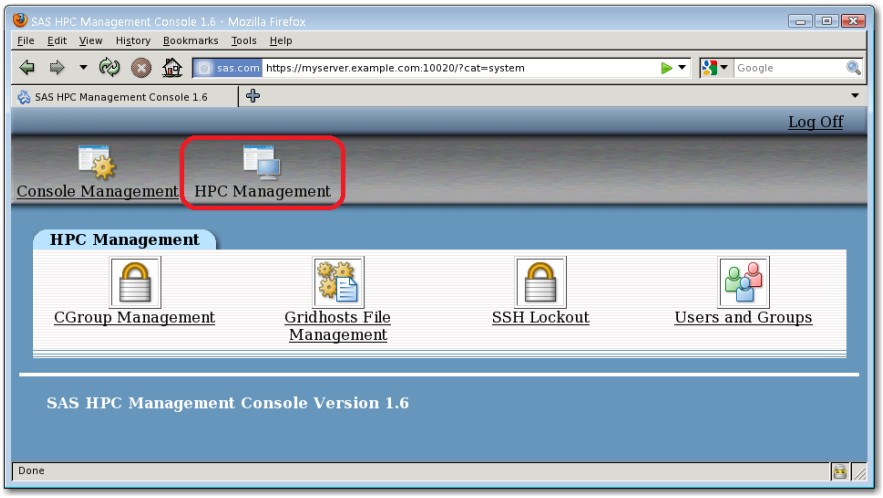
For example: **`http://myserver.example.com:10020`**

The Login page appears.



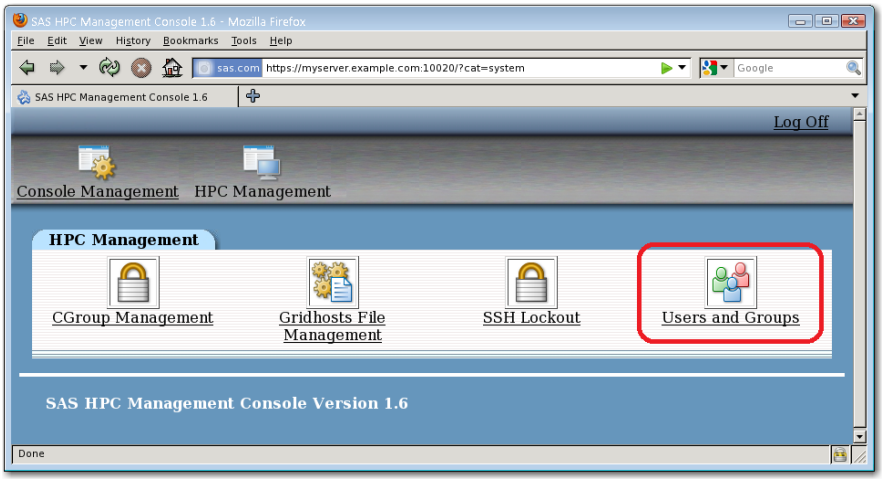
3 Log on to SAS High-Performance Computing Management Console.

The Console Management page appears.



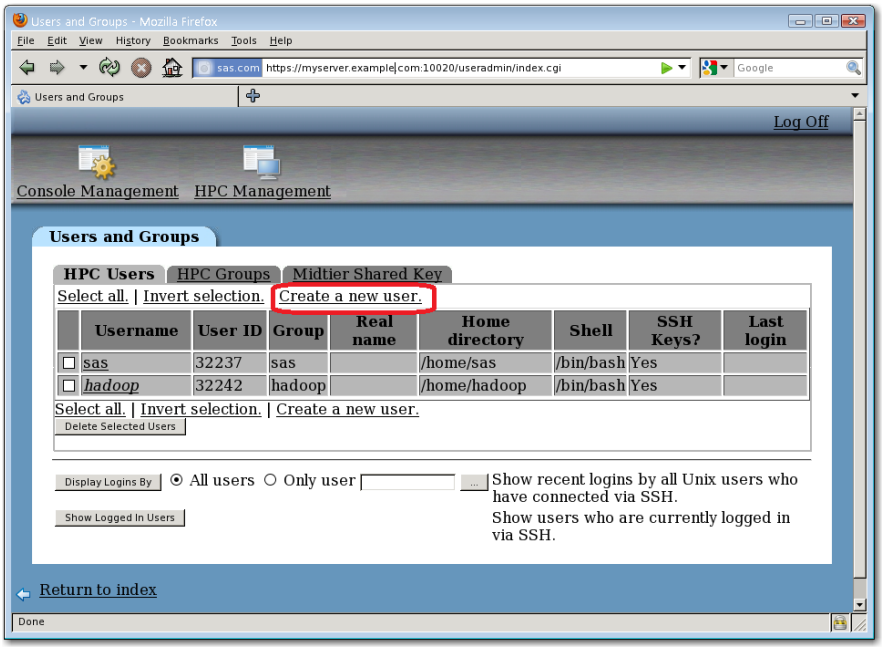
4 Click **HPC Management**.

The Console Management page appears.



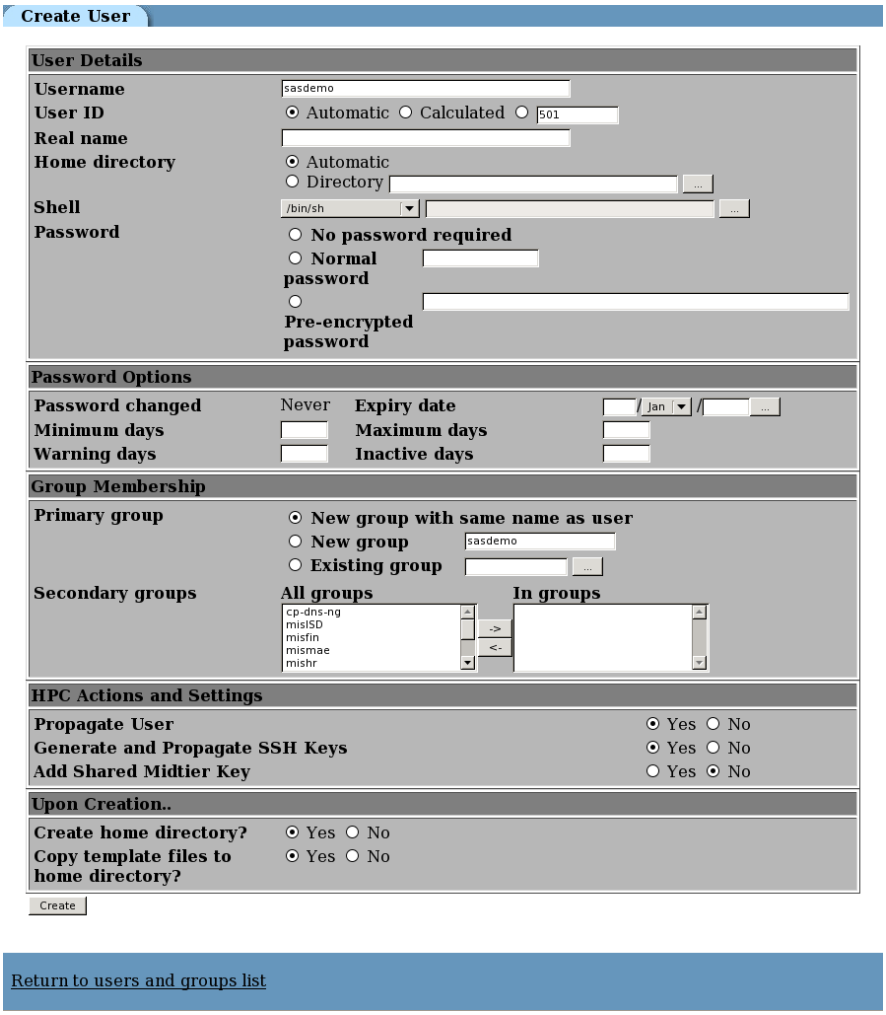
5 Click **Users and Groups**.

The Users and Groups page appears.



6 Click **Create a new user**.

The Create User page appears.



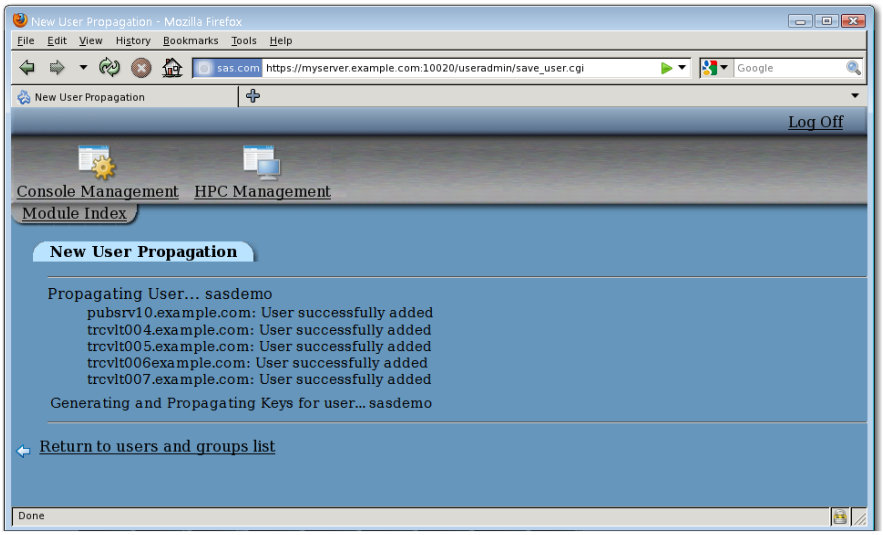
7 Enter information for the new user, using the security policies in place at your site.

Be sure to choose **Yes** for the following:

- **Propagate User**
- **Generate and Propagate SSH Keys**

When you are finished making your selections, click **Create**.

The New User Propagation page appears and lists the status of the create user command. Your task is successful if you see output similar to the following figure.


New User Propagation - Mozilla Firefox
File Edit View History Bookmarks Tools Help
https://myserver.example.com:10020/useradmin/save_user.cgi
New User Propagation
Log Off
Console Management
HPC Management
Module Index
New User Propagation
Propagating User... sasdemo
pubsrv10.example.com: User successfully added
trcvlt004.example.com: User successfully added
trcvlt005.example.com: User successfully added
trcvlt006example.com: User successfully added
trcvlt007.example.com: User successfully added
Generating and Propagating Keys for user... sasdemo
Return to users and groups list
Done

4

Installing and Configuring SAS High-Performance Deployment of Hadoop

SAS High-Performance Analytics Infrastructure Deployment Process Overview

Installing and configuring SAS High-Performance Deployment of Hadoop is the fifth of seven steps required to install and configure the SAS High-Performance Analytics infrastructure.

1. Review additional documentation.

2. Prepare your system.

3. Create a SAS Software Depot.

4. (Optional) Deploy SAS High-Performance Computing Management Console.

► **5. (Optional) Install and configure SAS High-Performance Deployment of Hadoop**

6. Deploy the SAS High-Performance Analytics environment.

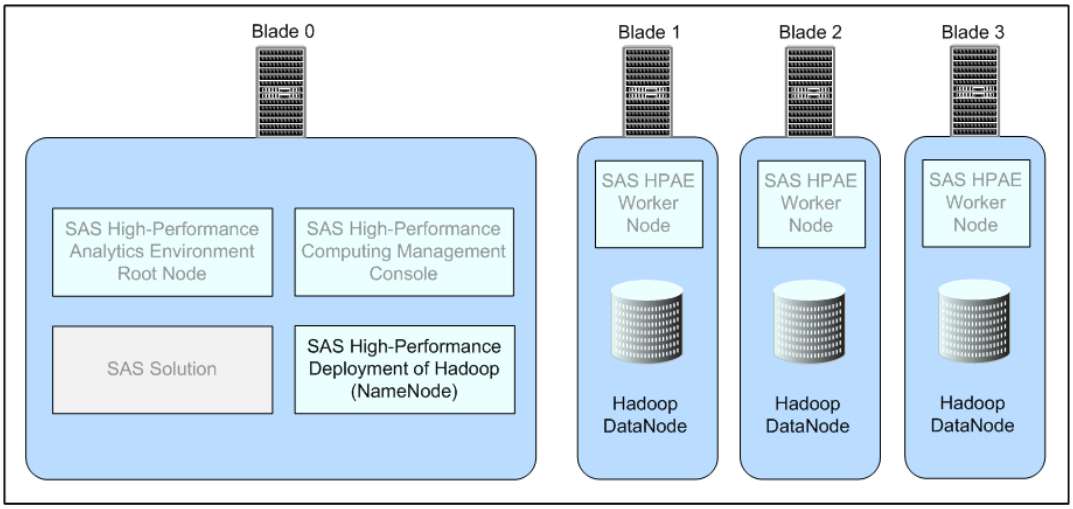
7. Configure your data storage.

Overview of Installing and Configuring SAS High-Performance Deployment of Hadoop

The SAS High-Performance Analytics environment relies on a massively parallel distributed database management system (Teradata or EMC Greenplum) or a Hadoop Distributed File System.

Deploying SAS High-Performance Deployment of Hadoop requires installing and configuring components on the NameNode machine and DataNodes on the remaining machines in the cluster. In this document, the NameNode is deployed on blade 0.

Figure 4.1 *Deploying the SAS High-Performance Deployment of Hadoop*



Install SAS High-Performance Deployment of Hadoop

The software that is needed for SAS High-Performance Deployment of Hadoop is available from within the SAS Software Depot that was created by the site depot administrator:

depot-installation-location`/standalone_installs/SAS_High_Performance_Hadoop_Deployment/2_1/Linux_for_x64/sashadoop.tar.gz`

On the machine designated as the Hadoop NameNode (blade 0), follow these steps:

1 Make sure that you have reviewed all of the information contained in the section "Preparing to Install and Configure SAS High-Performance Deployment of Hadoop" on page 15.

2 Log on to the Hadoop NameNode machine (blade 0) with a user account that has the necessary permissions.

For more information, see "User Accounts for SAS High-Performance Deployment of Hadoop" on page 15.

3 Decide where to install Hadoop, and create that directory if it does not exist.

```
mkdir hadoop
```

4 Record the name of this directory, as you will need it later in the install process.

5 Copy the sashadoop.tar.gz file to a temporary location and extract it:

```
cp sashadoop.tar.gz /tmp
cd /tmp
tar xzf sashadoop.tar.gz
```

A directory that is named **sashadoop** is created.

6 Change directory to the **sashadoop** directory and run the **hadoopInstall** command:

```
cd sashadoop
./hadoopInstall
```

7 Respond to the prompts from the configuration program:

Table 4.1 *SAS High-Performance Deployment of Hadoop Configuration Parameters*

Parameter	Description
Do you wish to use an existing Hadoop installation? (y/N)	Press Enter to perform a new installation.
Enter path to install Hadoop. The directory 'hadoop-0.23.1' will be created in the path specified.	Specify the directory that you created in Step 3 on page 42 and press Enter.
Enter replication factor. Default 2	Press Enter to accept the default or specify a preferred number of replications for blocks (0 - 10). This prompt corresponds to the **dfs.replication** property for HDFS.

Parameter	Description
Enter port number for fs.defaultFS. Default 54310 Enter port number for mapred.job.tracker. Default 54311 Enter port number for dfs.namenode.https-address. Default 50470 Enter port number for dfs.datanode.https.address. Default 50475 Enter port number for dfs.datanode.address. Default 50010 Enter port number for dfs.datanode.ipc.address. Default 50020 Enter port number for dfs.namenode.http-address. Default 50070 Enter port number for dfs.datanode.http.address. Default 50075 Enter port number for dfs.secondary.http.address. Default 50090 Enter port number for dfs.namenode.backup.address. Default 50100 Enter port number for dfs.namenode.backup.http-address. Default 50105 Enter port number for mapred.job.tracker.http.address. Default 50030 Enter port number for mapred.task.tracker.http.address. Default 50060 Enter port number for com.sas.lasr.hadoop.service.namenode.port. Default 15452 Enter port number for com.sas.lasr.hadoop.service.datanode.port. Default 15453	Press Enter for each prompt to accept the default port numbers. These ports are listed in “Pre-installation Ports Checklist for SAS” on page 21.

Parameter	Description
Enter user that will be running the HDFS server process.	Specify the user name and press Enter.
Enter path for JAVA_HOME directory. (Default: /usr/lib/jvm/jre)	Press Enter to accept the default JRE or specify the path to the JRE or JDK and press Enter. **Note:** The configuration program does not verify that a JRE is installed at **`/usr/lib/jvm/jre`**, that is the default path for some Linux vendors.
Enter path for Hadoop data directory. This should be on a large drive. Default is '/hadoop/hadoop-data'. Enter path for Hadoop system directory. Default is '/hadoop/hadoop-system'. Enter path for Hadoop local directory. Default is '/hadoop/hadoop-local'. Enter path for Hadoop name directory. Default is '/hadoop/hadoop-name'.	Press Enter to accept the default values or specify the paths that you prefer to use. **Note:** The data directory cannot be the root directory of a partition or mount. **Note:** If you have more than one data device, enter one of the data directories now, and after the installation, refer to "(Optional) Deploy with Multiple Data Devices" on page 45.
Enter full path to machine list. The NameNode '**`host`**' should be listed first.	Enter **`/etc/gridhosts`**.

8 The installation program installs SAS High-Performance Deployment of Hadoop on the local host, configures several files, and then provides a prompt:

```
The installer can now copy '/hadoop/hadoop-0.23.1' to all the slave
machines using scp, skipping the first entry. Perform copy?  (YES/no)
```

Enter **Yes** to install SAS High-Performance Deployment of Hadoop on the other machines in the cluster.

9 Choose the next step that applies to you:

- If you are using more than one data device, see "(Optional) Deploy with Multiple Data Devices" on page 45.

- If you want to limit the scope of sudo, see “(Optional) Limit the Scope of sudo” on page 46.
- If neither of the previous conditions apply to you, see “Format the Hadoop NameNode” on page 48.

(Optional) Deploy with Multiple Data Devices

If you plan to use more than one data device with the SAS High-Performance Deployment of Hadoop, then you must manually declare each device’s Hadoop data directory in hdfs-site.xml and push it out to all of your DataNodes.

To deploy SAS High-Performance Deployment for Hadoop with more than one data device, follow these steps:

1 Log on to the Hadoop NameNode using the account with which you plan to run Hadoop.

2 In a text editor, open ***hadoop-installation-directory*/etc/hadoop/hdfs-site.xml**.

3 Locate the `dfs.data.dir` property, specify the location of your additional data devices’ data directories, and save the file.

Separate multiple data directories with a comma. Enclose the directory list in double quotation marks.

For example:

```
dfs.data.dir = "/hadoop/hadoop-data,/data1/hadoop-data";
```

4 If you are planning to use the security wrapper, then refer to “(Optional) Limit the Scope of sudo” on page 46.

Otherwise, proceed to the next step.

5 Copy hdfs-site.xml to all of your Hadoop DataNodes using the **simcp** command.

For information about **simcp**, see Appendix 3, "SAS High-Performance Analytics Infrastructure Command Reference," on page 85.

6 Restart Hadoop with the following command:

```
HADOOP_HOME/sbin/start-dfs.sh
```

7 Proceed to "Format the Hadoop NameNode" on page 48.

(Optional) Limit the Scope of sudo

The SAS High-Performance Deployment for Hadoop requires that certain temporary directories be created on the NameNode and all of its DataNodes. To create these temporary directories—referred to as **sashdfstmp**—the system commands **mkdir** and **chown** are executed by Hadoop under the root user (through sudo). With the sample wrapper script provided in this document, you can confine use of **mkdir** and **chown** to the temporary directories.

To limit the scope of sudo, follow these steps:

1 Copy the sample wrapper script to a location on the Hadoop NameNode. You can copy the wrapper script from Appendix 5, "Sample Security Wrapper," on page 93.

2 Open the wrapper script in a text editor, locate the $tDir variable, specify the location of the Hadoop temporary data directory (**sashdfstmp**), and save the script.

Separate multiple temporary directories with a comma. Enclose the directory list in double quotation marks.

The location of **sashdfstmp** will always be at the root of the device (the mount point) where the Hadoop data directory resides. If you have multiple Hadoop data directories, then you must specify each location of the **sashdfstmp** directory with respect to each data directory's device.

For example:

```
my $tDir = "/sashdfstmp,/data1/sashdfstmp";
```

3 Copy the wrapper script to all of your Hadoop DataNodes using the **simcp** command. (The wrapper script must reside in the same location on every node.)

For information about **simcp**, see Appendix 3, "SAS High-Performance Analytics Infrastructure Command Reference," on page 85.

4 Make sure that you have properly modified your **/etc/sudoers** file. For more information, see "Modify the sudoers File" on page 16.

5 In a text editor, open ***hadoop-installation-directory*/etc/hadoop/hdfs-site.xml**.

6 Add lines similar to the following and save hdfs-site.xml:

```
<property>
  <name>com.sas.lasr.hadoop.security.script</name>
  <value>/hadoop/hadoop-0.23.1/sbin/my_security_wrapper.pl</value>
</property>
```

Note: If your site prefers to use an alternative command like **suexec** instead of sudo, set the command in the hdfs-site.xml file as follows:

```
<name>com.sas.lasr.hadoop.sudo.command</name>
  <value>/usr/sbin/suexec</value>
</property>
```

You can confirm that the command you specify is used by monitoring the NameNode log when you start Hadoop.

7 Copy hdfs-site.xml to all of your Hadoop DataNodes using the **simcp** command.

For information about **simcp**, see Appendix 3, "SAS High-Performance Analytics Infrastructure Command Reference," on page 85.

8 Restart Hadoop with the following command:

HADOOP_HOME/sbin/start-dfs.sh

When Hadoop starts, the temporary data directories are created.

9 Proceed to "Format the Hadoop NameNode" on page 48.

Format the Hadoop NameNode

To format the SAS High-Performance Deployment of Hadoop NameNode, follow these steps:

1 Make sure that you have modified the sudoers file on the Hadoop NameNode before proceeding.

See "Modify the sudoers File" on page 16.

2 Change to the **hadoop** user account:

```
su - hadoop
```

3 Export the HADOOP_HOME environment variable.

For example:

```
export "HADOOP_HOME=/hadoop/hadoop-0.23.1"
```

4 Format the NameNode:

```
hadoop-install-dir/hadoop-0.23.1/bin/hadoop namenode -format
```

5 At the `Re-format filesystem in /hadoop-install-dir/hadoop-name ? (Y or N)` prompt, enter **Y**. A line similar to the following highlighted output indicates that the format is successful:

```
Formatting using clusterid: CID-5b96061a-79f4-4264-87e0-99f351b749af
12/11/26 12:59:34 INFO util.HostsFileReader:
Refreshing hosts (include/exclude) list
12/11/26 12:59:35 INFO blockmanagement.DatanodeManager:
dfs.block.invalidate.limit=1000
12/11/26 12:59:35 INFO util.GSet: VM type       = 64-bit
12/11/26 12:59:35 INFO util.GSet: 2% max memory = 19.33375 MB
12/11/26 12:59:35 INFO util.GSet: capacity      = 2^21 = 2097152 entries
12/11/26 12:59:35 INFO util.GSet: recommended=2097152, actual=2097152
12/11/26 12:59:35 INFO blockmanagement.BlockManager:
dfs.block.access.token.enable=false
```

```
12/11/26 12:59:35 INFO blockmanagement.BlockManager: defaultReplication = 2
12/11/26 12:59:35 INFO blockmanagement.BlockManager: maxReplication     = 512
12/11/26 12:59:35 INFO blockmanagement.BlockManager: minReplication     = 1
12/11/26 12:59:35 INFO blockmanagement.BlockManager:
maxReplicationStreams      = 2
12/11/26 12:59:35 INFO blockmanagement.BlockManager:
shouldCheckForEnoughRacks  = false
12/11/26 12:59:35 INFO blockmanagement.BlockManager:
replicationRecheckInterval = 3000
12/11/26 12:59:35 INFO namenode.FSNamesystem: fsOwner=root (auth:SIMPLE)
12/11/26 12:59:35 INFO namenode.FSNamesystem: supergroup=supergroup
12/11/26 12:59:35 INFO namenode.FSNamesystem: isPermissionEnabled=true
12/11/26 12:59:35 INFO namenode.NameNode:
Caching file names occuring more than 10 times
12/11/26 12:59:36 INFO namenode.NNStorage: Storage directory
/hadoop/hadoop-name has been successfully formatted.
12/11/26 12:59:36 INFO namenode.FSImage: Saving image file
/hadoop/hadoop-name/current/fsimage.ckpt_0000000000000000000 using no compression
12/11/26 12:59:36 INFO namenode.FSImage: Image file of size 119 saved in 0 seconds.
12/11/26 12:59:36 INFO namenode.NNStorageRetentionManager:
Going to retain 1 images with txid >= 0
12/11/26 12:59:36 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at my_namenode.example.com/192.0.0.0
************************************************************/
```

6 While still using the **hadoop** user account, start the SAS High-Performance Deployment of Hadoop:

```
/hadoop-install-dir/hadoop-0.23.1/sbin/start-dfs.sh
```

A series of messages is printed to report the creation of log files and processes.

7 Create a directory in HDFS that permits Read and Write access for all users:

```
/hadoop-install-dir/hadoop-0.23.1/bin/hadoop fs -mkdir /hps
/hadoop-install-dir/hadoop-0.23.1/bin/hadoop fs -chmod 777 /hps
```

8 Proceed to "Validate SAS High-Performance Deployment of Hadoop" on page 50.

Validate SAS High-Performance Deployment of Hadoop

You can confirm that Hadoop is running successfully by opening a browser to **`http://`*`NameNode`*`:50070/dfshealth.jsp`**. Review the information in the cluster summary section of the page. Confirm that the number of live nodes equals the number of DataNodes and that the number of dead nodes is zero.

Note: It can take a few seconds for each node to start. If you do not see every node, then refresh the connection in the web interface.

5

Deploying the SAS High-Performance Analytics Environment

SAS High-Performance Analytics Infrastructure Deployment Process Overview

Installing and configuring the SAS High-Performance Analytics environment is the sixth of seven steps.

1. Review additional documentation.
2. Prepare your system.
3. Create a SAS Software Depot.
4. (Optional) Deploy SAS High-Performance Computing Management Console.
5. (Optional) Install and configure SAS High-Performance Deployment of Hadoop.
6. **Deploy the SAS High-Performance Analytics environment.**
7. Configure your data storage.

This chapter describes how to install and configure all of the components for the SAS High-Performance Analytics environment on the machines in the cluster.

Overview of Deploying the SAS High-Performance Analytics Environment

Deploying the SAS High-Performance Analytics environment requires installing and configuring components on the root node machine and on the remaining machines in the cluster. In this document, the root node is deployed on blade 0 (Hadoop), the Master Server (Greenplum), or the Teradata Managed Server (Teradata).

Figure 5.1 *SAS High-Performance Analytics Environment with a Data Appliance*

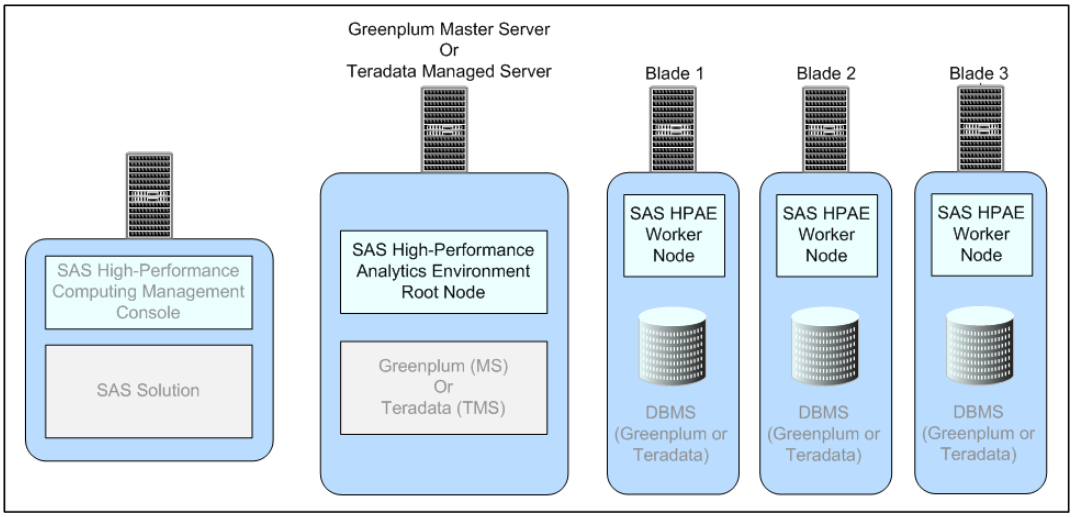
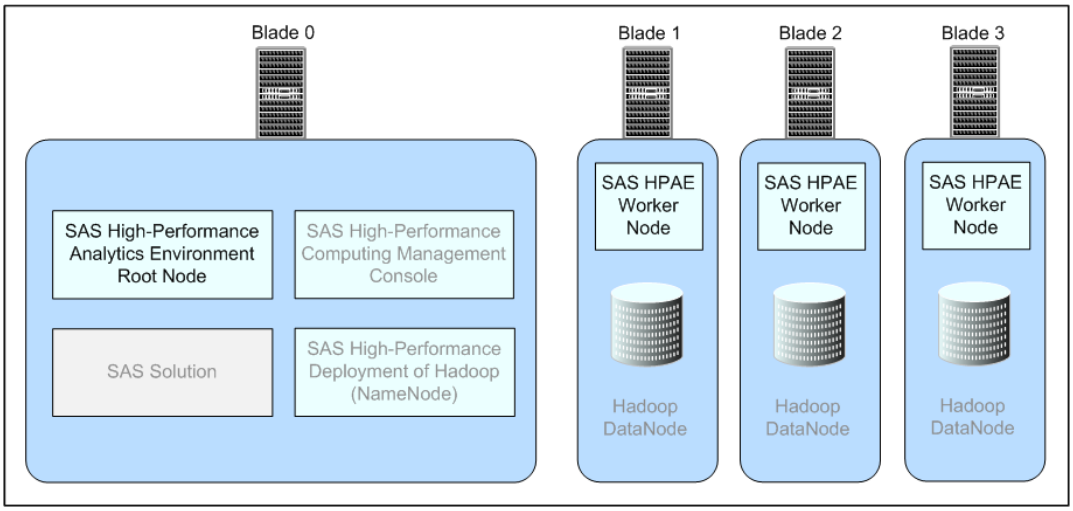


Figure 5.2 *SAS High-Performance Analytics Environment on a Hadoop Machine Cluster*



Install the SAS High-Performance Analytics Environment

The SAS High-Performance Analytics components can be installed with a shell script. Follow these steps to install with the script:

1 Make sure that you have reviewed all of the information contained in the section "Preparing to Install and Configure SAS High-Performance Deployment of Hadoop" on page 15.

2 The software that is needed for the SAS High-Performance Analytics environment is available from within the SAS Software Depot that was created by the site depot administrator: *depot-installation-location*`/standalone_installs/SAS_High-Performance_Node_Installation/2_1/Linux_for_x64`.

3 Copy the file TKGrid_Linux_x86_64.sh file to the `/tmp` directory of the root node of the cluster.

4 Log on to the machine that will serve as the root node of the cluster or the data appliance with a user account that has the necessary permissions.

   For more information, see "User Accounts for the SAS High-Performance Analytics Environment" on page 19.

5 Change directories to the desired installation location, such as `/opt`.

   Record the location of where you installed the analytics environment, as other configuration programs will prompt you for this path later in the deployment process.

6 Run the shell script in this directory.

   The shell script creates the `TKGrid` subdirectory and places all files under that directory.

7 Respond to the prompts from the configuration program:

Table 5.1 *Configuration Parameters*

Parameter	Description
Shared install or replicate to each node? (Y=SHARED/n=replicated)	If you are installing to a local drive on each node, then select **n** to indicate that this is a replicated installation. If you are installing to a drive that is shared across all the nodes (for example, NFS), then choose the shared installation.
Enter additional paths to include in LD_LIBRARY_PATH, separated by colons (:)	If you have any external library paths that you want to be accessible to the SAS High-Performance Analytics environment, enter the paths here.
Enter additional options to mpirun.	If you have any mpirun options to enter, do so here and press the Enter key. If you are using Kerberos, include the following option: `-genvlist `env | sed -e s/ =.*/,/ | sed /KRB5CCNAME/d | tr -d '\n'`TKPATH,LD_LIBRARY_PATH`
Enter path to use for Utility files. (default is /tmp).	SAS High-Performance Analytics applications might write scratch files. By default, these files are created in the **/tmp** directory. You can redirect the files to a different location by entering the path at the prompt. **Note:** If the directory that you specified does not exist, you must create it manually.
Enter path to Hadoop. (default is Hadoop not installed).	If your site uses Hadoop, enter the installation directory that you entered earlier in Step 3 on page 42. If your site does not use Hadoop, enter nothing and press the Enter or Return key.

Parameter	Description
Force Root Rank to run on headnode? (y/N)	If the appliance resides behind a firewall and only the root node can connect back to the client machines, select **y**. Otherwise, accept the default.
Enter full path to machine list...	Enter the name of the file that you created in the section "List the Machines in the Cluster or Appliance" (for example, **/etc/gridhosts**).
Enter maximum run time for grid jobs (in seconds). Default 7200 (2 hours).	If a SAS High-Performance Analytics application executes for more than the maximum allowable run time, it is automatically terminated. You can adjust that run-time limit here.
Enter value for UMASK. (default is unset.)	Enter a specific umask value and press the Enter key. Otherwise, simply press the Enter key. For more information, see "Consider Umask Settings" on page 20.

8 If you selected a replicated installation at the first prompt, you are now prompted to choose the technique for distributing the contents to the appliance nodes:

```
The install can now copy this directory to all the machines
listed in 'filename' using scp, skipping the first entry.
Perform copy? (YES/no)
```

Press Enter if you want the installation program to perform the replication. Enter **no** if you are distributing the contents of the installation directory by some other technique.

9 Proceed to "Validating the SAS High-Performance Analytics Environment Deployment" on page 57.

Validating the SAS High-Performance Analytics Environment Deployment

Overview of Validating the SAS High-Performance Analytics Environment Deployment

You have at least two methods to validate your SAS High-Performance Analytics environment deployment:

- “Use simsh to Validate the SAS High-Performance Analytics Deployment” on page 57.
- “Use MPI to Validate the SAS High-Performance Analytics Environment Deployment” on page 58.

Use simsh to Validate the SAS High-Performance Analytics Deployment

To validate your SAS High-Performance Analytics environment deployment by issuing a **`simsh`** command, follow these steps:

1. Log on to the machine where SAS High-Performance Computing Management Console is installed.
2. Enter the following command:

   `/HPCMC-install-dir/webmin/utilbin/simsh hostname`

   This command invokes the **`hostname`** command on each machine in the cluster. The host name for each machine is printed to the screen.

   You should see a list of known hosts similar to the following:

```
myblade006.example.com: myblade006.example.com
myblade007.example.com: myblade007.example.com
myblade004.example.com: myblade004.example.com
myblade005.example.com: myblade005.example.com
```

3 Proceed to “Configuring Your Data Storage” on page 60.

Use MPI to Validate the SAS High-Performance Analytics Environment Deployment

To validate your SAS High-Performance Analytics environment deployment by issuing a Message Passing Interface (MPI) command, follow these steps:

1 Log on to the root node using the SAS High-Performance Analytics environment installation account.

2 Enter the following command:

```
/HPA-environment-install-dir/TKGrid/mpich2-install/bin/mpirun
-f /etc/gridhosts hostname
```

You should see a list of known hosts similar to the following:

```
myblade006.example.com
myblade007.example.com
myblade004.example.com
myblade005.example.com
```

3 Proceed to “Configuring Your Data Storage” on page 60.

6

Configuring Your Data Storage

SAS High-Performance Analytics Infrastructure Deployment Process Overview

After you have deployed the SAS High-Performance Analytics infrastructure, you must configure your data storage to work with SAS.

Configuring your data storage is the last of seven deployment steps.

1. Review additional documentation.
2. Prepare your system.
3. Create a SAS Software Depot.
4. (Optional) Deploy SAS High-Performance Computing Management Console.
5. (Optional) Install and configure SAS High-Performance Deployment of Hadoop.
6. Deploy the SAS High-Performance Analytics environment.
7. **7. Configure your data storage.**

Overview of Configuring Your Data Storage

The SAS High-Performance Analytics environment relies on a massively parallel distributed database management system (Teradata or EMC Greenplum) or a Hadoop Distributed File System.

The topics that follow describe how you configure your site's data provider for the SAS High-Performance Analytics environment

Figure 6.1 *SAS High-Performance Analytics Infrastructure on a Supported Data Appliance*

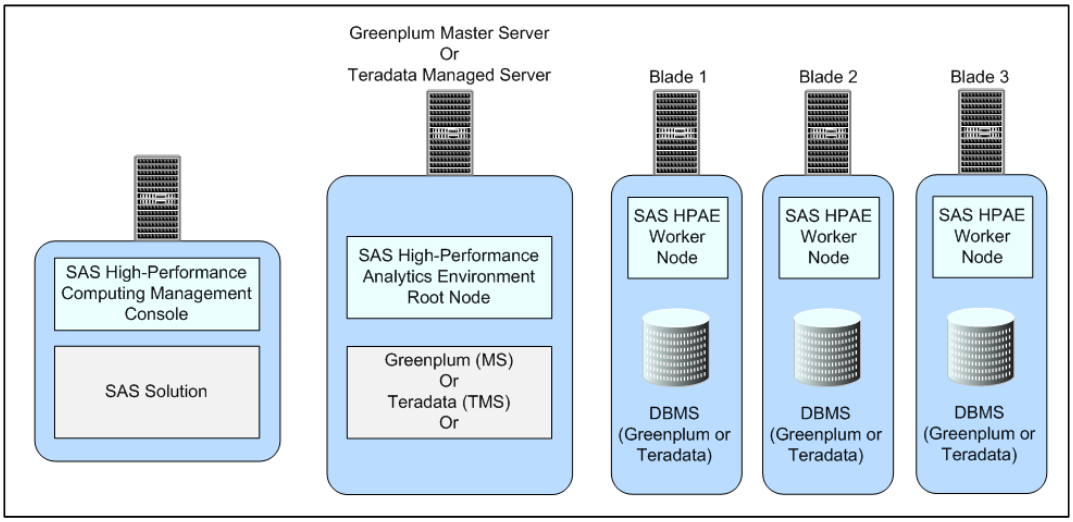
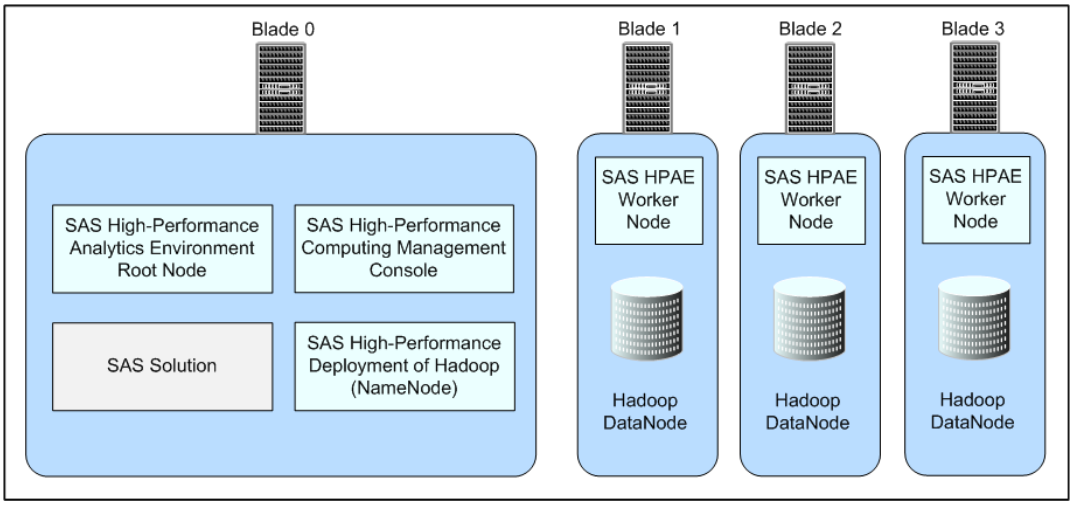


Figure 6.2 *SAS High-Performance Analytics Infrastructure on a Hadoop Machine Cluster*



Preparing the Greenplum Database for SAS Solutions

Configure the SAS/ACCESS Interface to Greenplum Software

SAS solutions, such as SAS High-Performance Analytics Server, rely on SAS/ACCESS to communicate with the Greenplum Data Appliance.

When you deploy the SAS/ACCESS Interface to Greenplum, make sure that the following configuration steps are performed:

1 On UNIX, make sure the following is done:

   a Set the ODBCHOME environment variable to your ODBC home directory.

   b Set the ODBCINI environment variable to the location and name of your odbc.ini file.

   TIP You can set both the ODBCHOME and ODBCINI environment variables in the SAS sasenv_local file and affect all executions of SAS. For more information, see *SAS Intelligence Platform: Data Administration Guide*, available at **http://support.sas.com//dcumentation/cdl/en/bidsag/65041/PDF/default/bidsag.pdf**.

   c Include the Greenplum ODBC drivers in your shared library path (LD_LIBRARY_PATH).

   d Edit odbc.ini and odbcinst.ini following the instructions listed in the *Configuration Guide for SAS Foundation for UNIX Environments*, available at **http://support.sas.com/documentation/installcenter/en/ikfdtnunxcg/66380/PDF/default/config.pdf**

   .

2 On Windows, make sure the following is done:

   a Configure a Data Source Name (DSN) to access the Greenplum database.

   b Register the SAS/ACCESS interface with the SAS system catalog.

   For more information about these Windows tasks, see the *Configuration Guide for SAS Foundation for Microsoft Windows for x64*, available at **http://support.sas.com/documentation/installcenter/en/ikfdtnwincg/66385/PDF/default/config.pdf**.

Create the SAS Protocol and Related Functions for Greenplum

The SAS High-Performance Analytics environment interface to the Greenplum database is supported by several functions that are associated with a dedicated SAS protocol.

To enable the functionality, the database administrator (a superuser such as the gpadmin role) should perform the following steps from the PostgreSQL environment or the PgAdmin III interface on the master server:

1 Install the formatter functions:

```
CREATE OR REPLACE FUNCTION formatter_export(record)
       RETURNS bytea
       AS '/HPA_environment-install-dir/TKGrid/lib/gpformatter.so',
'formatter_export'
       LANGUAGE C STABLE;
CREATE OR REPLACE FUNCTION formatter_import()
       RETURNS record
       AS '/HPA_environment-install-dir/TKGrid/lib/gpformatter.so',
       'formatter_import'
       LANGUAGE C STABLE;
```

2 Create the protocol functions and the SAS protocol:

```
CREATE OR REPLACE FUNCTION gpdb_to_sas() RETURNS integer
       AS '/HPA_environment-install-dir/TKGrid/lib/sas_gpext.so', 'sasprot_export'
       LANGUAGE C STABLE;
CREATE OR REPLACE FUNCTION sas_to_gpdb() RETURNS integer
       AS '/HPA_environment-install-dir/TKGrid/lib/sas_gpext.so', 'sasprot_import'
       LANGUAGE C STABLE;
```

```
CREATE TRUSTED PROTOCOL sas(readfunc='sas_to_gpdb', writefunc='gpdb_to_sas');
```

The functions must be created in a schema that is either in your schema search path or in the global pg_catalog catalog.

Each database role that executes SAS High-Performance Analytics code against the Greenplum database needs to be granted execution rights for the SAS protocol, as described in the next section.

3 These steps need to be repeated for each database that is accessed through SAS High-Performance Analytics procedures. You can define the functions on a database template from which new databases are derived.

Recommendations for Greenplum Database Roles

If multiple users access the SAS High-Performance Analytics environment on the Greenplum database, it is recommended that you set up a group role and associate the database roles for individual users with the group. The Greenplum database administrator can then associate access to the environment at the group level.

The following is one example of how you might accomplish this.

1 First, create the group.

For example:

```
CREATE GROUP sas_cust_group NOLOGIN;
ALTER ROLE sas_cust_group CREATEEXTTABLE;
```

Note: Remember that in Greenplum, only object privileges are inheritable. When granting the CREATEEXTTABLE, you are granting a system privilege. You can grant CREATEEXTTABLE to a group role, but the role must use a set role as a rolegroup first.

2 For each user, create a database role and associate it with the group.

For example:

```
CREATE ROLE megan  LOGIN IN ROLE sas_cust_group PASSWORD 'megan';
CREATE ROLE calvin LOGIN IN ROLE sas_cust_group PASSWORD 'calvin';
```

3 If a resource queue exists, associate the roles with the queue.

For example:

```
CREATE RESOURCE QUEUE sas_cust_queue WITH
                 (MIN_COST=10000.0   ,
                 ACTIVE_STATEMENTS=20,
                 PRIORITY=HIGH       ,
                 MEMORY_LIMIT='4GB'  );

ALTER ROLE megan  RESOURCE QUEUE sas_cust_queue;
ALTER ROLE calvin RESOURCE QUEUE sas_cust_queue;
```

4 Finally, grant the database roles execution rights on the SAS High-Performance Analytics protocol.

For example:

```
GRANT ALL ON PROTOCOL sas TO megan;
GRANT ALL ON PROTOCOL sas TO calvin;
```

Preparing the Teradata Database for the SAS High-Performance Analytics Environment

Configure the SAS/ACCESS Interface to Teradata

SAS solutions, such as SAS High-Performance Analytics Server, rely on SAS/ACCESS to communicate with the Teradata Managed Server Cabinet.

When you deploy the SAS/ACCESS Interface to Teradata, make sure that the following configuration steps are performed:

1 To perform FastExporting, the Teradata FastExport Utility must be present on the machine where you install SAS. You must also modify the library path (UNIX) and path (Windows) environment variables.

2 To perform MultiLoading, the Teradata MultiLoad Utility must be present on the system where you install SAS. You must also modify the path environment variable.

3 To use the Teradata parallel transporter API, the API must be installed on the machine where SAS is installed. You must also modify the path environment variable.

4 On UNIX, make sure the following is done:

   a Include the Teradata executable shared libraries in your shared library path.

   b HP-UX users must create two symbolic links.

   For more information about these UNIX tasks, see the *Configuration Guide for SAS 9.3 Foundation for UNIX Environments*, available at `http://support.sas.com/documentation/installcenter/en/ikfdtnunxcg/66380/PDF/default/config.pdf`.

5 On Windows, make sure that you verify connectivity by logging on to your Teradata account with the Teradata BTEQ utility.

   For more information, see the *Configuration Guide for SAS Foundation for Microsoft Windows for x64*, available at `http://support.sas.com/documentation/installcenter/en/ikfdtnwincg/66385/PDF/default/config.pdf`.

Install User-Defined Functions for Teradata

After deploying the SAS High-Performance Analytics environment, you must install the Teradata user-defined functions (UDFs) with the scripts provided in the RPM package.

On the Teradata Master Server (TMS), run the following command:

```
sh /HPA_environment-install-dir/TKGrid/bin/add_udfs.sh
```

You are prompted for a database account name with sufficient privileges to add user-defined functions.

Grant Privileges for Teradata Database Accounts

The database users who execute SAS High-Performance Analytics environment code on Teradata must have the following Teradata database privileges:

- `execute on SAS_SYSFNLIB`
- `select on SAS_SYSFNLIB`
- `execute function on SAS_SYSFNLIB`

Configure SAS High-Performance Deployment of Hadoop

Configuration of SAS High-Performance Deployment of Hadoop is performed at installation by the Hadoop deployment program, **`hadoopInstall`**.

Configure the Existing Cloudera Hadoop Cluster

Use the Cloudera Manager to configure your Cloudera 4 Hadoop deployment to interoperate with the SAS High-Performance Analytics environment.

1. Log on to the Cloudera Manager as an administrator.
2. If the DataNode service is present, stop it.
3. Remove the Remove Balancer service from the cluster.
4. Add SAS JAR files to the CDH4 library path. Copy both JAR files to **`/usr/lib/hadoop/lib`** (default location) on all nodes.

TIP You can issue a single **simcp** command to propagate JAR files across all machines in the cluster. The **simcp** and **simsh** commands are available with SAS High-Performance Computing Management Console. For more information, see Appendix 3, "SAS High-Performance Analytics Infrastructure Command Reference," on page 85.

5 Add the following to the plug-in configuration for the NameNode:

```
com.sas.lasr.hadoop.NameNodeService
```

6 Add the following to the plug-in configuration for DataNodes:

```
com.sas.lasr.hadoop.DataNodeService
```

7 Add the following lines to the advanced configuration for service-wide. These lines are placed in the HDFS Service Configuration Safety Valve property:

```
<property>
<name>com.sas.lasr.service.allow.put</name>
<value>true</value>
</property>
<property>
<name>com.sas.lasr.hadoop.service.namenode.port</name>
<value>15452</value>
</property>
<property>
<name>com.sas.lasr.hadoop.service.datanode.port</name>
<value>15453</value>
</property>
```

8 Add the following line to the advanced configuration for service-wide for the HDFS Service Environment Safety Valve:

```
"JAVA_HOME=/usr/lib/java/jdk1.7.0_07"
```

9 Restart all Cloudera Manager services.

10 Create and set the mode for the user directory in HDFS for testing by running the following commands as the Hadoop user. (Make sure that JAVA_HOME and HADOOP_HOME are set correctly before you run these commands.)

```
export JAVA_HOME=/usr/lib/java/jdk1.7.0_07
export HADOOP_HOME=/usr/lib/hadoop
```

11 Run the following commands to create the **/user** directory in HDFS:

```
$HADOOP_HOME/bin/hadoop fs -mkdir /user
$HADOOP_HOME/bin/hadoop fs -chmod 777 /user
```

12 Update and deploy the client configuration to each host in the cluster.

13 Add the following to the HDFS Client Configuration Safety Valve:

```
<property>
<name>com.sas.lasr.hadoop.service.namenode.port</name>
<value>15452</value>
</property>
<property>
<name>com.sas.lasr.hadoop.service.datanode.port</name>
<value>15453</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>file:///hadoop/hadoop-data</value>
</property>
```

14 Deploy the client configuration.

15 On the NameNode, in **/etc/hadoop/conf/hadoop-env.sh** add the following line:

```
"export JAVA_HOME=/usr/lib/java/jdk1.7.0_07"
```

16 Copy **/etc/hadoop/conf/hadoop-env.sh** to every machine in the cluster.

TIP You can issue a single **simcp** command to propagate hadoop-env.sh across all machines in the cluster. The **simcp** and **simsh** commands are available with SAS High-Performance Computing Management Console. For more information, see Appendix 3, “SAS High-Performance Analytics Infrastructure Command Reference,” on page 85.

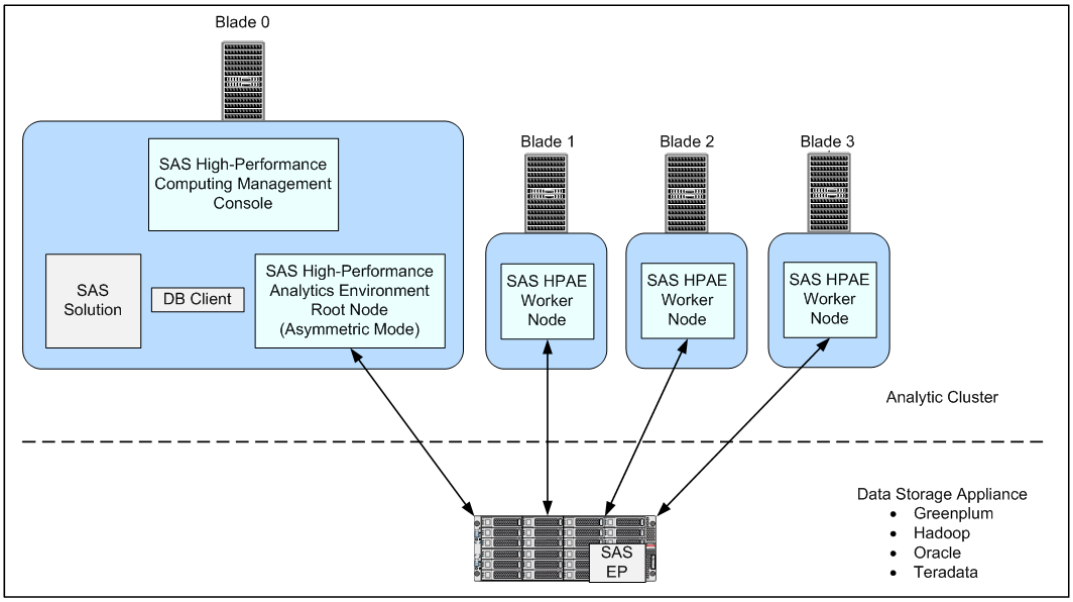
Appendix 1

Deploying the SAS High-Performance Analytics Environment in Asymmetric Mode

What is Asymmetric Mode?

Running the SAS High-Performance Analytics environment in asymmetric mode enables you to separate your data appliance or machine cluster from your analytics cluster. A SAS Embedded Process that resides on the data appliance is used to provide high-speed parallel data transfer between the data appliance and the analytics environment where it is processed.

Figure A1.1 *Topology for the SAS High-Performance Analytics Environment (Asymmetric Mode)*



Process Overview for Deploying in Asymmetric Mode

The process involved for installing and configuring the SAS High-Performance Analytics environment in asymmetric mode consists of the following steps:

1 Install and configure the SAS High-Performance Analytics environment on the analytics cluster.

For more information, see Chapter 5, “Deploying the SAS High-Performance Analytics Environment,” on page 51.

2 Gather information about the data provider that the analytics environment (in asymmetric mode) will query.

For more information, see “Overview of Preparing Your Data Provider” on page 73.

3 Install and configure the SAS High-Performance Analytics environment in *asymmetric mode* on the analytics cluster.

For more information, see “Deploy the SAS High-Performance Analytics Environment in Asymmetric Mode” on page 77.

Preparing Your Data Provider

Overview of Preparing Your Data Provider

Before you can install the SAS High-Performance Analytics environment in asymmetric mode, you must gather particular information about your data provider. If you are using Cloudera Hadoop, you must also complete a few configuration steps.

From the following list, choose the topic for your respective data provider:

1 “Prepare for Cloudera Hadoop” on page 73

2 “Prepare for a Greenplum Data Appliance” on page 75

3 “Prepare for an Oracle Exadata Appliance” on page 75

4 “Prepare for a Teradata Managed Server Cabinet” on page 76

Prepare for Cloudera Hadoop

Before you can install the SAS High-Performance Analytics environment in asymmetric mode to run with Cloudera Hadoop, there are certain requirements that must be met.

1 Copy the following Cloudera Hadoop JAR files into a directory on every machine in the analytics cluster:

- avro-1.5.4.jar
- commons-cli-1.2.jar

- commons-codec-1.4.jar
- commons-configuration-1.6.jar
- commons-httpclient-3.1.jar
- commons-lang-2.5.jar
- commons-logging-1.1.1.jar
- guava-11.0.2.jar
- hadoop-auth-2.0.0-cdh4.0.1.jar
- hadoop-common-2.0.0-cdh4.0.1.jar
- hadoop-core-2.0.0-mr1-cdh4.0.1.jar
- hadoop-hdfs-2.0.0-cdh4.0.1.jar
- jackson-core-asl-1.8.8.jar
- jackson-mapper-asl-1.8.8.jar
- jsp-api-2.1.jar
- log4j-1.2.16.jar
- protobuf-java-2.4.0a.jar
- slf4j-api-1.6.1.jar
- slf4j-log4j12-1.6.1.jar

2 Record the path to the Cloudera Hadoop JAR files required by SAS in the table that follows:

Table A1.1 *Record the Location of the Cloudera Hadoop JAR Files Required by SAS*

Example	Actual Path of the Required Cloudera Hadoop JAR Files on Your System
/opt/TKGrid_REP/Cloudera	

3 Record the path to the 64-bit Java Runtime Engine (JRE) required by the SAS Threaded Kernel JNI (version 1.5 or later) in the table that follows:

Table A1.2 *Record the Location of the JRE*

Example	Actual Path of the JRE on Your System
/opt/java/jre1.7.0_07	

Prepare for a Greenplum Data Appliance

Before you can install the SAS High-Performance Analytics environment in asymmetric mode to run with a Greenplum data appliance, there are certain requirements that must be met.

1 Install the Greenplum client on every machine in your analytics cluster.

For more information, refer to your Greenplum documentation.

2 Record the path to the Greenplum client in the table that follows:

Table A1.3 *Record the Location of the Greenplum Client*

Example	Actual Path of the Greenplum Client on Your System
/usr/local/greenplum-db	

Prepare for an Oracle Exadata Appliance

Before you can install the SAS High-Performance Analytics environment in asymmetric mode to run with an Oracle Exadata appliance, there are certain requirements that must be met.

1 Install the Oracle client on every machine in your analytics cluster.

For more information, refer to your Oracle documentation.

2 Record the path to the Oracle client in the table that follows. (This should be the absolute path to libclntsh.so):

Table A1.4 *Record the Location of the Oracle Client*

Example	Actual Path of the Oracle Client on Your System
/usr/local/ora11gr2/product/11.2.0/client_1/lib	

3 Record the value of the Oracle TNS_ADMIN environment variable in the table that follows. (Typically, this is the directory that contains the tnsnames.ora file):

Table A1.5 *Record the Value of the Oracle TNS_ADMIN Environment Variable*

Example	Oracle TNS_ADMIN Environment Variable Value on Your System
/my_server/oracle	

Prepare for a Teradata Managed Server Cabinet

Before you can install the SAS High-Performance Analytics environment in asymmetric mode to run with a Teradata Managed Server Cabinet, there are certain requirements that must be met.

1 Install the Teradata client on every machine in your analytics cluster.

For more information, refer to your Teradata documentation.

2 Record the path to the Teradata client in the table that follows. (This should be the absolute path to the directory that contains the `odbc_64` subdirectory):

Table A1.6 Record the Location of the Teradata Client

Example	Actual Location of the Teradata Client on Your System
/opt/teradata/client/13.10	

Deploy the SAS High-Performance Analytics Environment in Asymmetric Mode

The SAS High-Performance Analytics environment in asymmetric mode is deployed using a shell script. Follow these steps to install with the script:

1 Make sure that you have reviewed all of the information contained in the section "Preparing Your Data Provider" on page 73.

2 The software that is needed for the SAS High-Performance Analytics environment is available from within the SAS Software Depot that was created by the site depot administrator: *depot-installation-location*`/standalone_installs/SAS_High-Performance_Node_Installation/2_1/Linux_for_x64`.

3 Copy the file TKGrid_Linux_REP_x86_64.sh to the `/tmp` directory of the root node of the analytic cluster.

4 Log on to the machine that will serve as the root node of the cluster with a user account that has the necessary permissions.

For more information, see "User Accounts for the SAS High-Performance Analytics Environment" on page 19.

5 Change directories to the desired installation location, such as `/opt`.

6 Run the shell script in this directory.

The shell script creates the **TKGrid_REP** subdirectory and places all files under that directory.

7 Respond to the prompts from the configuration program:

Table A1.7 Configuration Parameters

Parameter	Description
Do you want to configure remote access to Teradata? (yes/NO)	If you are using a Teradata Managed Cabinet for your data provider, enter **y** and press Enter. Otherwise, enter **n** and press Enter.
Do you want to use Teradata client installed in /opt/teradata/client/13.10 ? (YES/no)	If you have installed the Teradata client in the default path, then enter nothing and press Enter. Otherwise, enter **n** and press Enter.
Enter path of Teradata client install. i.e.: /opt/teradata/client/13.10	If you chose **n** in the previous step, enter the path where the Teradata client was installed. (This path was recorded earlier in Table A1.6 on page 77.)
Do you want to configure remote access to Greenplum? (yes/NO)	If you are using a Greenplum Data Appliance for your data provider, enter **y** and press Enter. Otherwise, enter **n** and press Enter.
Do you want to use Greenplum client installed in /usr/local/greenplum-db ? (YES/no)	If you have installed the Greenplum client in the default path, then enter nothing and press Enter. Otherwise, enter **n** and press Enter.
Enter path of Greenplum client install. i.e.: /usr/local/greenplum-db	If you chose **n** in the previous step, enter the path where the Greenplum client was installed. (This path was recorded earlier in Table A1.3 on page 75.)
Do you want to configure remote access to Hadoop? (yes/NO)	If you are using a Cloudera Hadoop machine cluster for your data provider, enter **y** and press Enter. Otherwise, enter **n** and press Enter.

Parameter	Description
Do you want to use the JRE installed in /opt/java/jre1.7.0_07 ?	If you want to use the JRE at the path that the install program lists, then enter nothing and press Enter. Otherwise, enter **n** and press Enter.
Enter path of the JRE i.e.: /opt/java/jre1.7.0_07	If you chose **n** in the previous step, enter the path where the JRE was installed. (This path was recorded earlier in Table A1.2 on page 75.)
Enter path of the directory containing the Hadoop and SAS/EP jars.	Enter the path where the Cloudera Hadoop JAR files required by SAS reside. (This path was recorded earlier in Table A1.1 on page 74.)
Do you want to configure remote access to Oracle? (yes/NO)	If you are using an ORACLE Exadata appliance for your data provider, enter **y** and press Enter. Otherwise, enter **n** and press Enter.
Enter path of Oracle client libraries. i.e.: /usr/local/ora11gr2/product/11.2.0/client_1/lib	Enter the path where the Oracle client libraries reside. (This path was recorded earlier in Table A1.4 on page 76.)
Enter path of TNS_ADMIN, or just enter if not needed.	Enter the value of the Oracle TNS_ADMIN environment variable. (This value was recorded earlier in Table A1.5 on page 76.)
Shared install or replicate to each node? (Y=SHARED/n=replicated)	If you are installing to a local drive on each node, then select **n** to indicate that this is a replicated installation. If you are installing to a drive that is shared across all the nodes (for example, NFS), then choose the shared installation.

Parameter	Description
Enter path to TKGrid install	Enter the absolute path to where the SAS High-Performance Analytics environment is installed. This should be the directory in which the analytics environment install program was run with **TKGrid** appended to it (for example, **/opt/TKGrid**). For more information, see Step 5 on page 54.
Enter additional paths to include in LD_LIBRARY_PATH, separated by colons (:)	If you have any external library paths that you want to be accessible to the SAS High-Performance Analytics environment, enter the paths here.

8 If you selected a replicated installation at the first prompt, you are now prompted to choose the technique for distributing the contents to the appliance nodes:

```
The install can now copy this directory to all the machines
listed in 'pathname' using scp, skipping the first entry. Perform copy?  (YES/no)
```

Press Enter if you want the installation program to perform the replication. Enter **no** if you are distributing the contents of the installation directory by some other technique.

9 You have finished deploying the SAS High-Performance Analytics environment in asymmetric mode. If you have not done so already, install the appropriate SAS Embedded Process on the *data* appliance or *data* machine cluster for your respective data provider.

For more information, see *SAS In-Database Products: Administrator's Guide*, available at **http://support.sas.com/documentation/cdl/en/indbag/66665/PDF/default/indbag.pdf**.

Appendix 2

Updating the SAS High-Performance Analytics Infrastructure

Overview of Updating the SAS High-Performance Analytics Infrastructure

Here are some considerations for updating the SAS High-Performance Analytics infrastructure:

- Because of dependencies, if you update the SAS High-Performance Analytics environment, you must also update SAS High-Performance Deployment of Hadoop.
- Update SAS High-Performance Deployment of Hadoop first, followed by the SAS High-Performance Analytics environment.

Update SAS High-Performance Computing Management Console

To update your deployment of SAS High-Performance Computing Management Console, follow these steps:

1 Stop the server by entering the following command:

```
service sashpcmc stop
```

2 Update the management console using the following RPM command:

```
rpm -U /SAS-Software-Depot-Root-Dir/standalone_installs/
SAS_High-Performance_Management_Console/2_1/Linux_for_x64/sashpcmc-1.6.x86_64.rpm
```

Update SAS High-Performance Deployment of Hadoop

To update SAS High-Performance Deployment of Hadoop, follow these steps:

1 Stop SAS High-Performance Deployment of Hadoop by running the **`/hadoop/hadoop/sbin/stop-all.sh`** command with the **`hadoop`** user account on the NameNode before you perform any action.

2 Check that there are no Java processes owned by **`hadoop`** running on any machine:

```
ps -ef | grep hadoop
```

If you find any Java processes owned by the **`hadoop`** user account, terminate them.

TIP If you have High-Performance Computing Management Console installed, you can issue a single **`simsh`** command to simultaneously check all the machines

in the cluster: `/HPCMC-install-dir/webmin/utilbin/simsh ps -ef | grep hadoop`.

3 Delete the Hadoop installation directory on every machine in the cluster:

```
rm -r -f /hadoop-install-dir
```

TIP If you have High-Performance Computing Management Console installed, you can issue a single **simsh** command to simultaneously remove the Hadoop install directories on all the machines in the cluster: `/HPCMC-install-dir/webmin/utilbin/simsh rm -r -f /hadoop-install-dir`.

4 Re-install Hadoop using **hadoopInstall** as described in "Install SAS High-Performance Deployment of Hadoop" on page 41.

5 Use the **hadoop** user account to run the **/hadoop/hadoop/sbin/start-all.sh** command on the NameNode.

Confirm that SAS High-Performance Deployment of Hadoop is running successfully by opening a browser to **http://*namenode*:50070/dfshealth.jsp**. Review the information in the cluster summary section of the page. Confirm that the number of live nodes equals the number of DataNodes and that the number of dead nodes is zero.

Update the SAS High-Performance Analytics Environment

Updating your deployment of the SAS High-Performance Analytics environment consists of deleting the deployment and reinstalling the newer version. To update the SAS High-Performance Analytics environment, follow these steps:

1 Check that there are no SAS High-Performance Analytics environment processes running on any machine:

```
ps -ef | grep TKGrid
```

If you find any TKGrid processes, terminate them.

TIP If you have High-Performance Computing Management Console installed, you can issue a single **simsh** command to simultaneously check all the machines in the cluster: `/HPCMC-install-dir/webmin/utilbin/simsh ps –ef | grep TKGrid`.

2 Delete the SAS High-Performance Analytics environment installation directory on every machine in the cluster:

```
rm -r -f /HPA-environment-install-dir
```

TIP If you have High-Performance Computing Management Console installed, you can issue a single **simsh** command to simultaneously remove the environment install directories on all the machines in the cluster: `/HPCMC-install-dir/webmin/utilbin/simsh rm -r -f /HPA-environment-install-dir`.

3 Re-install the SAS High-Performance Analytics environment using the shell script as described in “Install the SAS High-Performance Analytics Environment” on page 54.

Appendix 3

SAS High-Performance Analytics Infrastructure Command Reference

The **simsh** and **simcp** commands are installed with SAS High-Performance Computing Management Console. The default path to the commands is ***/HPCMC-install-dir/webmin/utilbin***. Any user account that can access the commands and has passwordless secure shell configured can use them.

The **simsh** command uses secure shell to invoke the specified command on every machine that is listed in the **/etc/gridhosts** file. The following command demonstrates invoking the **hostname** command on each machine in the cluster:

```
/HPCMC-install-dir/webmin/utilbin/simsh hostname
```

TIP You can use SAS High-Performance Computing Management Console to create and manage your grid hosts file. For more information, see *SAS High-Performance Computing Management Console: User's Guide*, available at **http://support.sas.com/documentation/onlinedoc/va/index.html**.

The **simcp** command is used to copy a file from one machine to the other machines in the cluster. Passwordless secure shell and an **/etc/gridhosts** file are required. The following command is an example of copying the **/etc/hosts** file to each machine in the cluster:

```
/HPCMC-install-dir/webmin/utilbin/simcp /etc/hosts /etc
```

Appendix 4

Deploying on SELinux and IPTables

Overview of Deploying on SELinux and IPTables

This document describes how to prepare Security Enhanced Linux (SELinux) and IPTables for a SAS High-Performance Analytics infrastructure deployment.

Security Enhanced Linux (SELinux) is a feature in some versions of Linux that provides a mechanism for supporting access control security policies. IPTables is a firewall—a combination of a packet-filtering framework and generic table structure for defining rulesets. SELinux and IPTables is available in most new distributions of Linux, both community-based and enterprise-ready. For sites that require added security, the use of SELinux and IPTables is an accepted approach for many IT departments.

Because of the limitless configuration possibilities, this document is based on the default configuration for SELinux and IPTables running on RedHat Enterprise Linux (RHEL) 6.3. You might need to adjust the directions accordingly, especially for complex SELinux and IPTables configurations.

Prepare SAS High-Performance Computing Management Console

SELinux Modifications for SAS High-Performance Computing Management Console

After generating and propagating root's SSH keys throughout the cluster or data appliance, you must run the following command on every machine or blade to restore the security context on the files in **`/root/.ssh`**:

```
restorecon -R -v /root/.ssh
```

IPTables Modifications for SAS High-Performance Computing Management Console

Add the following line to **/etc/sysconfig/iptables** to allow connections to the port on which the management console is listening (10020 by default). Open the port only on the machine on which the management console is running:

```
-A INPUT -m state --state NEW -m tcp -p tcp --dport 10020 -j ACCEPT
```

Prepare SAS High-Performance Deployment of Hadoop

SELinux Modifications for SAS High-Performance Deployment of Hadoop

After generating and propagating root's SSH keys throughout the cluster or data appliance, you must run the following command on every machine or blade to restore the security context on the files in **/root/.ssh**:

```
restorecon -R -v /root/.ssh
```

IPTables Modifications for SAS High-Performance Deployment of Hadoop

Hadoop has a number of ports on which it communicates. To open these ports, place the following lines in **/etc/sysconfig/iptables**:

Note: The following example uses default ports. Modify as necessary for your site.

```
-A INPUT -m state --state NEW -m tcp -p tcp --dport 54310 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 54311 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 50470 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 50475 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 50010 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 50020 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 50070 -j ACCEPT
```

```
-A INPUT -m state --state NEW -m tcp -p tcp --dport 50075 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 50090 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 50100 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 50105 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 50030 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 50060 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 15452 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 15453 -j ACCEPT
```

Edit **`/etc/sysconfig/iptables`** and then copy this file across the machine cluster or data appliance. Lastly, restart the IPTables service.

Prepare SAS High-Performance Analytics Environment

SELinux Modifications for SAS High-Performance Analytics Environment

After generating and propagating root's SSH keys throughout the cluster or data appliance, you must run the following command on every machine or blade to restore the security context on the files in **`/root/.ssh`**:

```
restorecon -R -v /root/.ssh
```

IPTables Modifications for SAS High-Performance Analytics Environment

If you are deploying the SAS LASR Analytic Server, then you must define one port per server in **`/etc/sysconfig/iptables`**. (The port number is defined in the SAS code that starts the LASR server.)

If you have more than one server running simultaneously, you need all these ports defined in the form of a range.

The following is an example of an iptables entry for a single server (one port):

```
-A INPUT -m state --state NEW -m tcp -p tcp --dport 10010 -j ACCEPT
```

The following is an example of an iptables entry for five servers (port range):

```
-A INPUT -m state --state NEW -m tcp -p tcp --dport 10010:10014 -j ACCEPT
```

MPICH_PORT_RANGE must also be opened in IPTables by editing the **/etc/sysconfig/iptables** file and adding the port range.

The following is an example for five servers:

```
-A INPUT -m state --state NEW -m tcp -p tcp --dport 10010:10029 -j ACCEPT
```

Edit **/etc/sysconfig/iptables** and then copy this file across the machine cluster or data appliance. Lastly, restart the IPTables service.

SAS High-Performance Analytics Environment Post-Installation Modifications

The SAS High-Performance Analytics environment uses Message Passing Interface (MPI) communications, which requires you to define one port range per active job across the machine cluster or data appliance.

(A port range consists of a minimum of four ports per active job. Every running monitoring server counts as a job on the cluster or appliance.)

For example, if you have five jobs running simultaneously across the machine cluster or data appliance, you need a minimum of 20 ports in the range.

The following example is an entry in tkmpirsh.sh for five jobs:

```
export MPICH_PORT_RANGE=18401:18420
```

Edit tkmpirsh.sh using the number of jobs appropriate for your site. (tkmpirsh.sh is located in ***/installation-directory*/TKGrid/**.) Then, copy tkmpirsh.sh across the machine cluster or data appliance.

iptables File

This topic lists the complete **/etc/sysconfig/iptables** file. The additions to iptables described in this document are highlighted.

```
*filter
:INPUT ACCEPT [0:0]
:FORWARD ACCEPT [0:0]
:OUTPUT ACCEPT [0:0]
-A INPUT -m state --state ESTABLISHED,RELATED -j ACCEPT
-A INPUT -p icmp -j ACCEPT
-A INPUT -i lo -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 22 -j ACCEPT
# Needed by SAS HPC MC
-A INPUT -m state --state NEW -m tcp -p tcp --dport 10020 -j ACCEPT
# Needed for HDFS (Hadoop)
A INPUT -m state --state NEW -m tcp -p tcp --dport 54310 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 54311 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 50470 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 50475 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 50010 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 50020 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 50070 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 50075 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 50090 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 50100 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 50105 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 50030 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 50060 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 15452 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 15453 -j ACCEPT
# End of HDFS Additions
# Needed for LASR Server Ports.
-A INPUT -m state --state NEW -m tcp -p tcp --dport 17401:17405 -j ACCEPT
# End of LASR Additions
# Needed for MPICH.
-A INPUT -m state --state NEW -m tcp -p tcp --dport 18401:18420 -j ACCEPT
# End of MPICH additions.
-A INPUT -j REJECT --reject-with icmp-host-prohibited
-A FORWARD -j REJECT --reject-with icmp-host-prohibited
```

Appendix 5

Sample Security Wrapper

The SAS High-Performance Deployment of Hadoop requires elevated privileges for temporary directories only. The following Perl script is a sample wrapper that you can modify and copy to your SAS High-Performance for Hadoop system. The purpose of this wrapper is to confine the system commands available to the Hadoop user through sudo. For more information, see "(Optional) Limit the Scope of sudo" on page 46.

```
#!/usr/bin/perl
# Name: sm_wrapper.pl
#
# Description: This script acts as a wrapper to the system
# commands mkdir and chown, which are executed as root via
# sudo by the user running HDFS. The script is written to
# confine operations to the temp directories used by LASR.
# The script should be the only command the HDFS user executes
# via sudo. In addition to a proper sudoers file entry, the
# property com.sas.lasr.hadoop.security.script must be defined
# in hdfs-site.xml with a valid path to the script has the
# value.
#
# Error Code Descriptions:
#
# 200: Unauthorized directory was used as an argument to the
# command
#
# 201: Error in mkdir function. Something prevented the
# creation of the directory, like the directory already
# exists or file permissions.
#
# 202: Error in getting user information for chown. The user
# does not exist or nsswitch.conf does not have the entries
# required to resolve the user.
```

```
#
# 203: Error in chown function. Something prevented the
# ownership changed, like the directory does not exist.
#
# 210: Attempt at using ".."
#

use strict;
use File::Spec;

###
# Beginning of user defined strings
###

# Location of sashdfstmp directory
#
# The location of the sashdfstmp will always be at the root
# of the device where the hadoop data directory resides. If you
# have multiple hadoop data directories, then you must specify
# each location of the sashdfstmp directory with respect to
# each data directory's device.
#
# Example: "/sashdfstmp"
#        or
#        "/sashdfstmp,/data1/sashdfstmp

my $tDir = "/sashdfstmp,/data1/sashdfstmp";

###
# End of user defined strings
# DO NOT EDIT BELOW THIS LINE
###

my $cmd;

$cmd = shift @ARGV;
chomp($cmd);

if ( $cmd eq 'mkdir' ) {
    my $retCode = 0;
    my $dir = pop @ARGV;
    chomp($dir);
    my $tRetC = &validate_Path($dir, $tDir);
    if ( $tRetC > 0 ) {
        if ($tRetC == 200) {
            print "Unauthorized directory for $cmd: Error Code $tRetC\n";
            exit $tRetC;
```

```
        }
        if ($tRetC == 210) {
            print "Unauthorized attempt at reverse directory traversal:
            Error Code $tRetC\n";
            exit $tRetC;
        }
    }
    mkdir($dir, 0000) or $retCode = 201;
    if ($retCode == 0) {
        exit 0;
    } else {
        print "Error in mkdir for $dir: Error Code $retCode\n";
        exit $retCode;
    }
} elsif ( $cmd eq 'chown' ) {
    my $retCode = 0;
        my $cUser = shift @ARGV;
    chomp($cUser);
        my $dir = shift @ARGV;
    chomp($dir);
    my $tRetC = &validate_Path($dir, $tDir);
    if ( $tRetC > 0 ) {
        if ($tRetC == 200) {
            print "Unauthorized directory for $cmd: Error Code $tRetC\n";
            exit $tRetC;
        }
        if ($tRetC == 210) {
            print "Unauthorized attempt at reverse directory traversal:
            Error Code $tRetC\n";
            exit $tRetC;
        }
    }
        $cUser =~ s/://g;
    my ($login,$pass,$uid,$gid) = getpwnam($cUser) or $retCode = 202;
    chown($uid,$gid,$dir) or $retCode = 203;
    if ($retCode == 0) {
        exit 0;
    } elsif ($retCode == 202) {
        print "Error in getting user info for $cUser: Error Code $retCode\n";
        exit $retCode;
    } elsif ($retCode == 203) {
        print "Error in chown of $dir: Error Code $retCode\n";
        exit $retCode;
    }
} else {
    print "Unauthorized command used as argument to the wrapper script\n";
    exit;
```

```
}

sub validate_Path {
    my $vDir = shift @_;
    my $rPath = shift @_;
    my @vDirs = File::Spec->splitdir($vDir);
    my $vDirLen = scalar @vDirs;
    my $dirMatch = 0;
    my $tCheck = 0;

    if ( $rPath =~ /,/ ) {
        my @dDirC = split(",", $rPath);
        foreach my $tPath (@dDirC) {
            my @sDirs = File::Spec->splitdir($tPath);
            my $sDirLen = scalar @sDirs;
            $dirMatch = 0;
            $tCheck = 0;
            my $c =0;
            for ($c; $c < $sDirLen; $c++) {
                if ($vDirs[$c] eq $sDirs[$c]) {
                    $dirMatch = 1;
                } else {
                    $dirMatch = 0;
                    if ($vDirs[$c] =~ /\.\./) {
                        $tCheck = 1;
                        last;
                    }
                }
            }
            if ($dirMatch == 1) {
                for ($c; $c < $vDirLen; $c++) {
                    if ($vDirs[$c] =~ /\.\./) {
                        $tCheck = 1;
                        last;
                    }
                }
                last;
            }
            if ($tCheck == 1) {
                last;
            }
        }
    } else {
        my @sDirs = File::Spec->splitdir($rPath);
        my $sDirLen = scalar @sDirs;
        $dirMatch = 0;
        $tCheck = 0;
```

```
        my $c =0;
        for ($c; $c < $sDirLen; $c++) {
            if ($vDirs[$c] eq $sDirs[$c]) {
                $dirMatch = 1;
            } else {
                $dirMatch = 0;
                if ($vDirs[$c] =~ /\.\./) {
                    $tCheck = 1;
                    last;
                }
            }
        }
        if ($dirMatch == 1) {
            for ($c; $c < $vDirLen; $c++) {
                if ($vDirs[$c] =~ /\.\./) {
                    $tCheck = 1;
                    last;
                }
            }
        }
    }

    if ($tCheck == 1) {
        return 210;
    } elsif ($dirMatch == 0) {
        return 200;
    } else {
        return 0;
    }
}
```

Glossary

data set
See SAS data set

encryption
the act or process of converting data to a form that is unintelligible except to the intended recipients.

foundation services
See SAS Foundation Services

grid host
the machine to which the SAS client makes an initial connection in a SAS High-Performance Analytics application.

Hadoop Distributed File System
a framework for managing files as blocks of equal size, which are replicated across the machines in a Hadoop cluster to provide fault tolerance.

HDFS
See Hadoop Distributed File System

identity
See metadata identity

Integrated Windows authentication
a Microsoft technology that facilitates use of authentication protocols such as Kerberos. In the SAS implementation, all participating components must be in the same Windows domain or in domains that trust each other.

Internet Protocol Version 6
See IPv6

IPv6
a protocol that specifies the format for network addresses for all computers that are connected to the Internet. This protocol, which is the successor of Internet Protocol Version 4, uses hexadecimal notation to represent 128-bit address spaces. The format can consist of up to eight groups of four hexadecimal characters, delimited by colons, as in FE80:0000:0000:0000:0202:B3FF:FE1E:8329. As an alternative, a group of consecutive zeros could be replaced with two colons, as in FE80::0202:B3FF:FE1E:8329. Short form: IPv6

IWA
See Integrated Windows authentication

JAR file
a Java Archive file. The JAR file format is used for aggregating many files into one file. JAR files have the file extension .jar.

Java
a set of technologies for creating software programs in both stand-alone environments and networked environments, and for running those programs safely. Java is an Oracle Corporation trademark.

Java Database Connectivity
See JDBC

Java Development Kit
See JDK

JDBC
a standard interface for accessing SQL databases. JDBC provides uniform access to a wide range of relational databases. It also provides a common base on which higher-level tools and interfaces can be built. Short form: JDBC.

JDK
a software development environment that is available from Oracle Corporation. The JDK includes a Java Runtime Environment (JRE), a compiler, a debugger, and other tools for developing Java applets and applications. Short form: JDK.

localhost
the keyword that is used to specify the machine on which a program is executing. If a client specifies localhost as the server address, the client connects to a server that runs on the same machine.

login
a SAS copy of information about an external account. Each login includes a user ID and belongs to one SAS user or group. Most logins do not include a password.

Message Passing Interface
is a message-passing library interface specification. SAS High-Performance Analytics applications implement MPI for use in high-performance computing environments.

metadata identity
a metadata object that represents an individual user or a group of users in a SAS metadata environment. Each individual and group that accesses secured resources on a SAS Metadata Server should have a unique metadata identity within that server.

metadata object
a set of attributes that describe a table, a server, a user, or another resource on a network. The specific attributes that a metadata object includes vary depending on which metadata model is being used.

middle tier
in a SAS business intelligence system, the architectural layer in which Web applications and related services execute. The middle tier receives user requests, applies business logic and business rules, interacts with processing servers and data servers, and returns information to users.

MPI

See Message Passing Interface

object spawner

a program that instantiates object servers that are using an IOM bridge connection. The object spawner listens for incoming client requests for IOM services. When the spawner receives a request from a new client, it launches an instance of an IOM server to fulfill the request. Depending on which incoming TCP/IP port the request was made on, the spawner either invokes the administrator interface or processes a request for a UUID (Universal Unique Identifier).

planned deployment

a method of installing and configuring a SAS business intelligence system. This method requires a deployment plan that contains information about the different hosts that are included in the system and the software and SAS servers that are to be deployed on each host. The deployment plan then serves as input to the SAS Deployment Wizard.

root node

in a SAS High-Performance Analytics application, the role of the software that distributes and coordinates the workload of the worker nodes. In most deployments the root node runs on the machine that is identified as the grid host. SAS High-Performance Analytics applications assign the highest MPI rank to the root node.

SAS Application Server

a logical entity that represents the SAS server tier, which in turn comprises servers that execute code for particular tasks and metadata objects.

SAS authentication

a form of authentication in which the target SAS server is responsible for requesting or performing the authentication check. SAS servers usually meet this responsibility by asking another component (such as the server's host operating system, an LDAP provider, or the SAS Metadata Server) to perform the check. In a few cases (such as SAS internal authentication to the metadata server), the SAS server performs the check for itself. A configuration in which a SAS server trusts that another component

has pre-authenticated users (for example, Web authentication) is not part of SAS authentication.

SAS configuration directory
the location where configuration information for a SAS deployment is stored. The configuration directory contains configuration files, logs, scripts, repository files, and other items for the SAS software that is installed on the machine.

SAS data set
a file whose contents are in one of the native SAS file formats. There are two types of SAS data sets: SAS data files and SAS data views.

SAS Deployment Manager
a cross-platform utility that manages SAS deployments. The SAS Deployment Manager supports functions such as updating passwords for your SAS deployment, rebuilding SAS Web applications, and removing configurations.

SAS Deployment Wizard
a cross-platform utility that installs and initially configures many SAS products. Using a SAS installation data file and, when appropriate, a deployment plan for its initial input, the wizard prompts the customer for other necessary input at the start of the session, so that there is no need to monitor the entire deployment.

SAS Foundation Services
a set of core infrastructure services that programmers can use in developing distributed applications that are integrated with the SAS platform. These services provide basic underlying functions that are common to many applications. These functions include making client connections to SAS application servers, dynamic service discovery, user authentication, profile management, session context management, metadata and content repository access, activity logging, event management, information publishing, and stored process execution.

SAS installation data file
See SID file

SAS installation directory
the location where your SAS software is installed. This location is the parent directory to the installation directories of all SAS products. The SAS installation directory is also referred to as SAS Home in the SAS Deployment Wizard.

SAS IOM workspace
in the IOM object hierarchy for a SAS Workspace Server, an object that represents a single session in SAS.

SAS Metadata Server
a multi-user server that enables users to read metadata from or write metadata to one or more SAS Metadata Repositories.

SAS Pooled Workspace Server
a SAS Workspace Server that is configured to use server-side pooling. In this configuration, the SAS object spawner maintains a collection of workspace server processes that are available for clients.

SAS Software Depot
a file system that consists of a collection of SAS installation files that represents one or more orders. The depot is organized in a specific format that is meaningful to the SAS Deployment Wizard, which is the tool that is used to install and initially configure SAS. The depot contains the SAS Deployment Wizard executable, one or more deployment plans, a SAS installation data file, order data, and product data.

SAS Stored Process Server
a SAS IOM server that is launched in order to fulfill client requests for SAS Stored Processes.

SAS Workspace Server
a SAS IOM server that is launched in order to fulfill client requests for IOM workspaces.

SASHDAT file
the data format used for tables that are added to HDFS by SAS. SASHDAT files are read in parallel by the server.

SASHOME directory
the file location where an instance of SAS software is installed on a computer. The location of the SASHOME directory is established at the initial installation of SAS software by the SAS Deployment Wizard. That location becomes the default installation location for any other SAS software you install on the same machine.

server context
a SAS IOM server concept that describes how SAS Application Servers manage client requests. A SAS Application Server has an awareness (or context) of how it is being used and makes decisions based on that awareness. For example, when a SAS Data Integration Studio client submits code to its SAS Application Server, the server determines what type of code is submitted and directs it to the correct physical server for processing (in this case, a SAS Workspace Server).

server description file
a file that is created by a SAS client when the LASR procedure executes to create a server. The file contains information about the machines that are used by the server. It also contains the name of the server signature file that controls access to the server.

SID file
a control file containing license information that is required in order to install SAS.

spawner
See object spawner

worker node
in a SAS High-Performance Analytics application, the role of the software that receives the workload from the root node.

workspace

See SAS IOM workspace

Index

O

P

R

S

T