SAS® High-Performance Analytics Infrastructure 2.5

Installation and Configuration Guide

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2014. *SAS® High-Performance Analytics Infrastructure 2.5: Installation and Configuration Guide*. Cary, NC: SAS Institute Inc.

SAS® High-Performance Analytics Infrastructure 2.5: Installation and Configuration Guide

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

Electronic book 1, March 2014

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

Accessibility

For information about the accessibility of any of the products mentioned in this document, see the usage documentation for that product.

x

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/ACCESS for Relational Databases: Reference*, `http://support.sas.com/documentation/onlinedoc/access/index.html`.
- *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 In-Database Products: Administrator's Guide*, available at `http://support.sas.com/documentation/onlinedoc/indbtech/index.html`.
- *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 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, and optionally, SAS High-Performance Computing Management Console, and SAS High-Performance Deployment of 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.

Which Version Do I Use?

This document is published for each major release of the SAS High-Performance Analytics infrastructure, which consists of the following products:

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

  (referred to as the SAS High-Performance Node Installation)

Refer to your order summary to determine the specific version of the infrastructure that is included in your SAS order. Your order summary resides in your SAS Software Depot

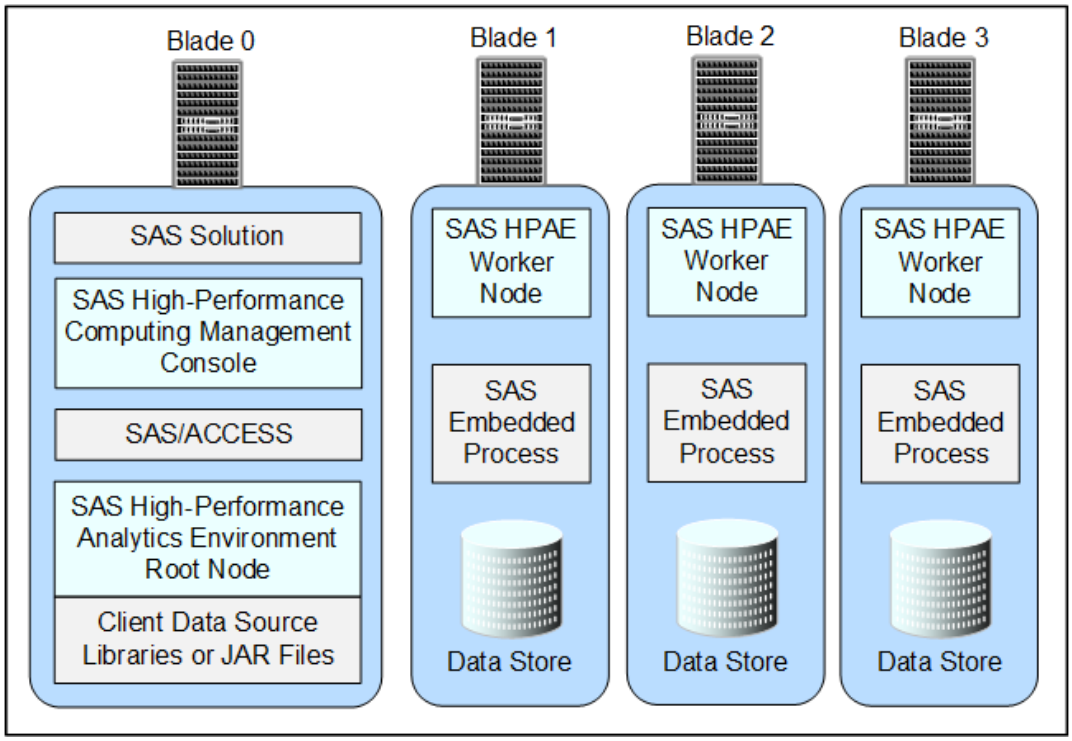
for your respective order under the **`install_doc`** directory (for example, **`C:\SAS Software Depot\install_doc\my-order\ordersummary.html`**).

What is the Infrastructure?

The SAS High-Performance Analytics infrastructure consists of software that performs analytic tasks in a high-performance environment, which is characterized by massively parallel processing (MPP). The infrastructure is used by SAS products and solutions that typically analyze big data that resides in a distributed data storage appliance or Hadoop cluster.

The following figure depicts the SAS High-Performance Analytics infrastructure in its most basic topology:

Figure 1.1 *SAS High-Performance Analytics Infrastructure Topology (Simplified)*



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

- SAS High-Performance Analytics environment

  The SAS High-Performance Analytics environment is the core of the infrastructure. The environment performs analytic computations on an analytic cluster. The analytics cluster is a Hadoop cluster or a data appliance.

- (Optional) SAS High-Performance Deployment of Hadoop

  Some solutions such as SAS Visual Analytics rely on a SAS data store that is co-located with the SAS High-Performance Analytic environment on the analytic cluster. The SAS High-Performance Deployment for Hadoop provides a Hadoop implementation that is pre-configured for use with the SAS High-Performance Analytics environment. Alternatively, these solutions can use a pre-existing Hadoop deployment or one of the supported data appliances.

- (Optional) SAS High-Performance Computing Management Console

  The SAS High-Performance Computing Management Console is used to ease the administration of distributed, high-performance computing (HPC) environments. Tasks such as configuring passwordless SSH, propagating user accounts and public keys, and managing CPU and memory resources on the analytic cluster are all made easier by the management console.

Other software on the analytics cluster include the following:

- SAS/ACCESS Interface and SAS Embedded Process

  Together the SAS/ACCESS Interface and SAS Embedded Process provide a high-speed parallel connection that delivers data from the co-located SAS data source to the SAS-High Performance Analytics environment on the analytic cluster. These components are contained in a deployment package that is specific for your data source.

  For more information, refer to the *SAS In-Database Products: Administrator's Guide*, available at **http://support.sas.com/documentation/onlinedoc/indbtech/index.html** and the *SAS/ACCESS for Relational Databases: Reference* available at **http://support.sas.com/documentation/onlinedoc/access/index.html**.

Note: For deployments that use Hadoop for the co-located data provider and access SASHDAT tables exclusively, SAS/ACCESS and SAS Embedded Process is not needed.

- Database client libraries or JAR files

  Data vendor-supplied client libraries—or in the case of Hadoop, JAR files—are required for the SAS Embedded Process to transfer data to and from the data store and the SAS High-Performance Analytics environment.

- SAS solutions

  The SAS High-Performance Analytics infrastructure is used by various SAS High-Performance solutions such as the following:

  - SAS High-Performance Analytics Server

    For more information, refer to `http://support.sas.com/documentation/onlinedoc/securedoc/index_hpa.html`.

  - SAS High-Performance Marketing Optimization

    For more information, refer to `http://support.sas.com/documentation/onlinedoc/mktopt/index.html`.

  - SAS High-Performance Risk

    For more information, refer to `http://support.sas.com/documentation/onlinedoc/hprisk/index.html`.

  - SAS Visual Analytics

    For more information, refer to `http://support.sas.com/documentation/onlinedoc/va/index.html`.

Where Do I Locate My Analytics Cluster?

Overview of Locating Your Analytics Cluster

You have two options for where to locate your SAS analytics cluster:

- Co-locate SAS with your data store.
- Separate SAS from your data store.

  When your SAS analytics cluster is separated (remote) from your data store, you have two basic options for transferring data:

  - Serial data transfer using SAS/ACCESS.
  - Parallel data transfer using SAS/ACCESS in conjunction with the SAS Embedded Process.

The topics in this section contain simple diagrams that describe each option for analytic cluster placement:

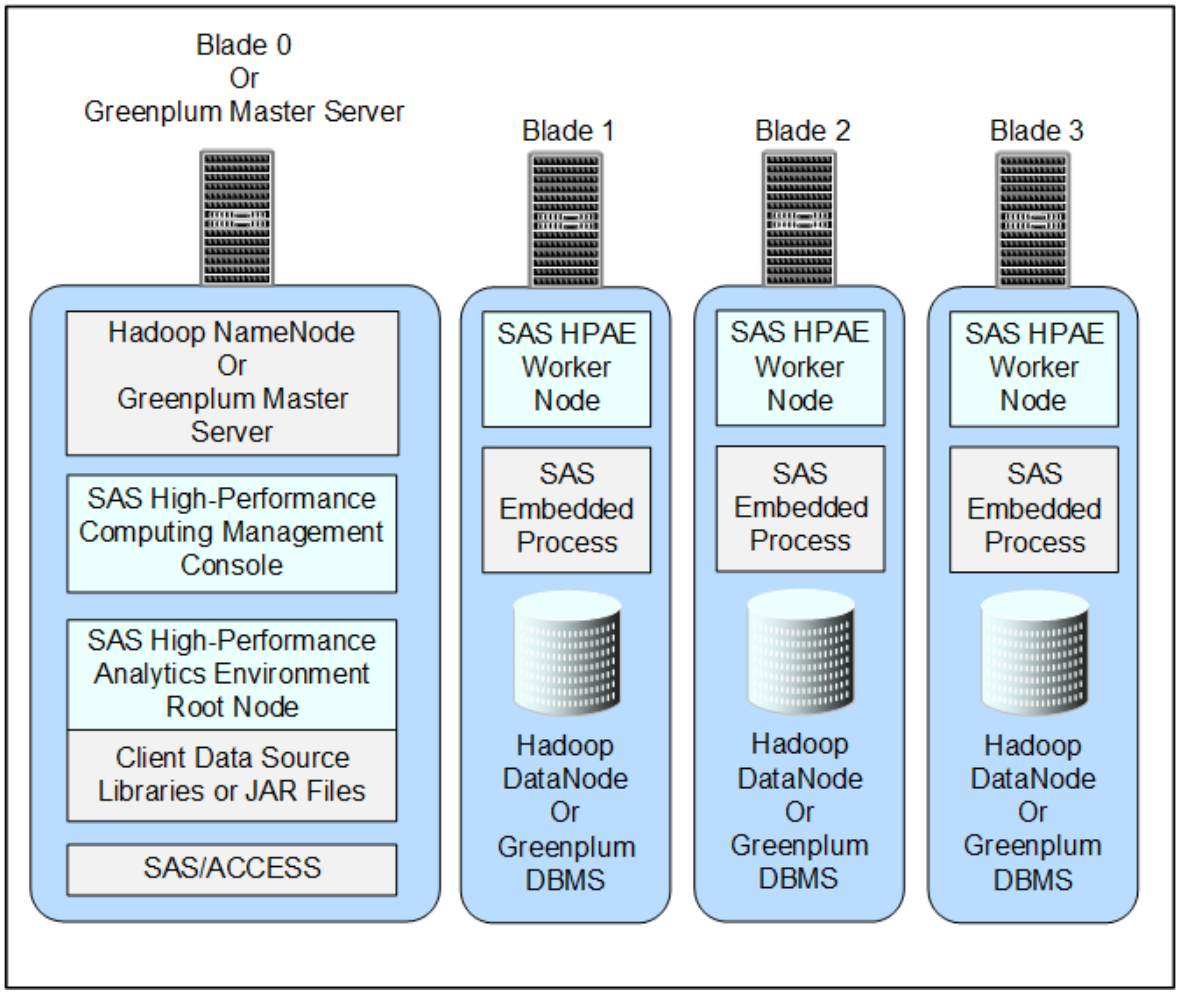
- Co-Located with the data store
- Remote from the data store (serial connection)
- Remote from the data store (parallel connection)

TIP Where you locate your cluster depends on a number of criteria. Your SAS representative will know the latest supported configurations, and can work with you to help you determine which cluster placement option works best for your site. Also, there might be solution-specific criteria that you should consider when determining your analytics cluster location. For more information, see the installation or administration guide for your specific SAS solution.

Analytic Cluster Co-Located with Your Data Store

The following figure shows the analytics cluster co-located on your Hadoop cluster or Greenplum data appliance:

Figure 1.2 *Analytics Cluster Co-Located on the Hadoop Cluster or Greenplum Data Appliance*

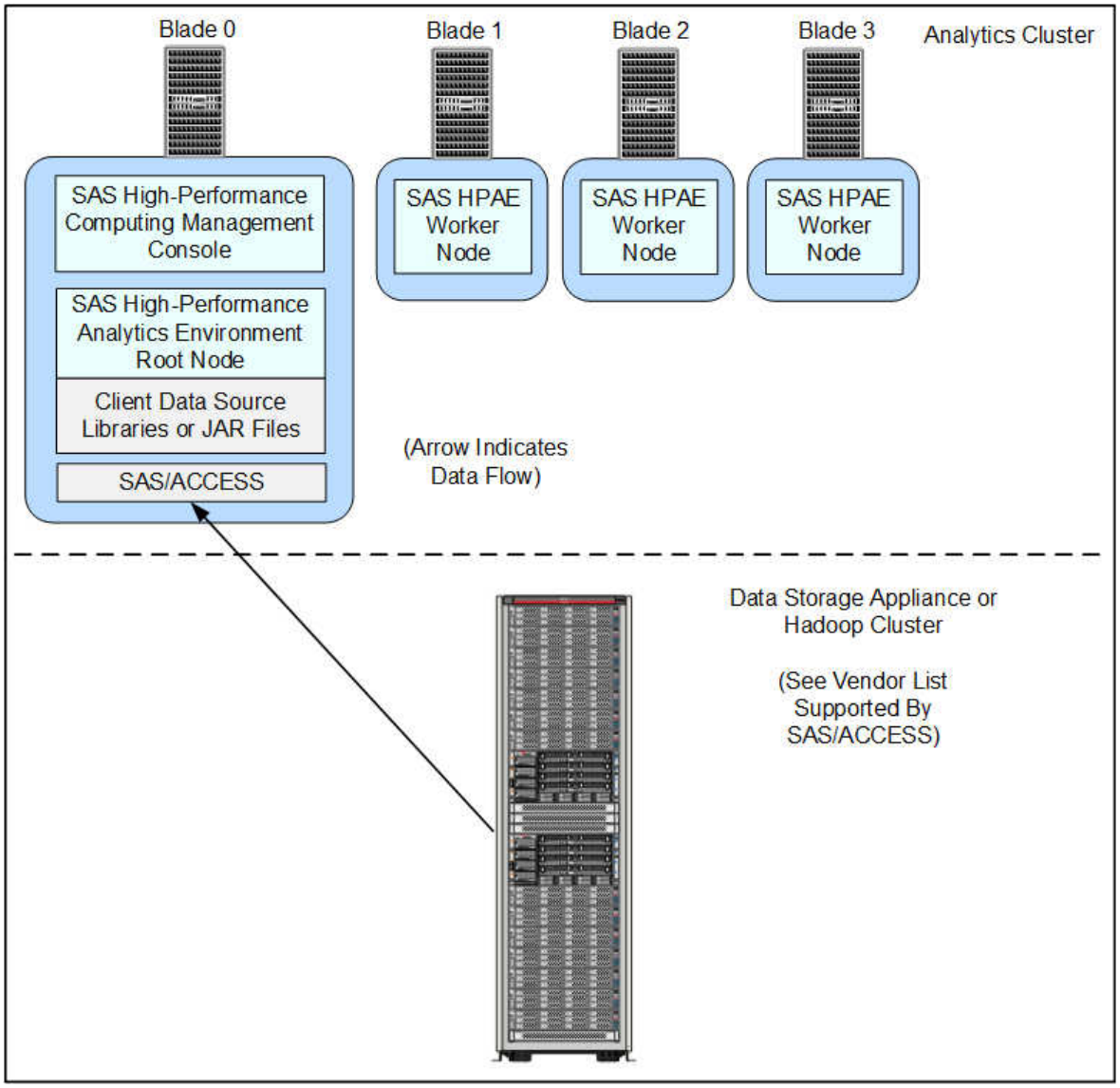


Note: For deployments that use Hadoop for the co-located data provider and access SASHDAT tables exclusively, SAS/ACCESS and the SAS Embedded Process are not needed.

Analytic Cluster Remote from Your Data Store (Serial Connection)

The following figure shows the analytics cluster using a serial connection to your remote data store:

Figure 1.3 *Analytics Cluster Remote from Your Data Store (Serial Connection)*

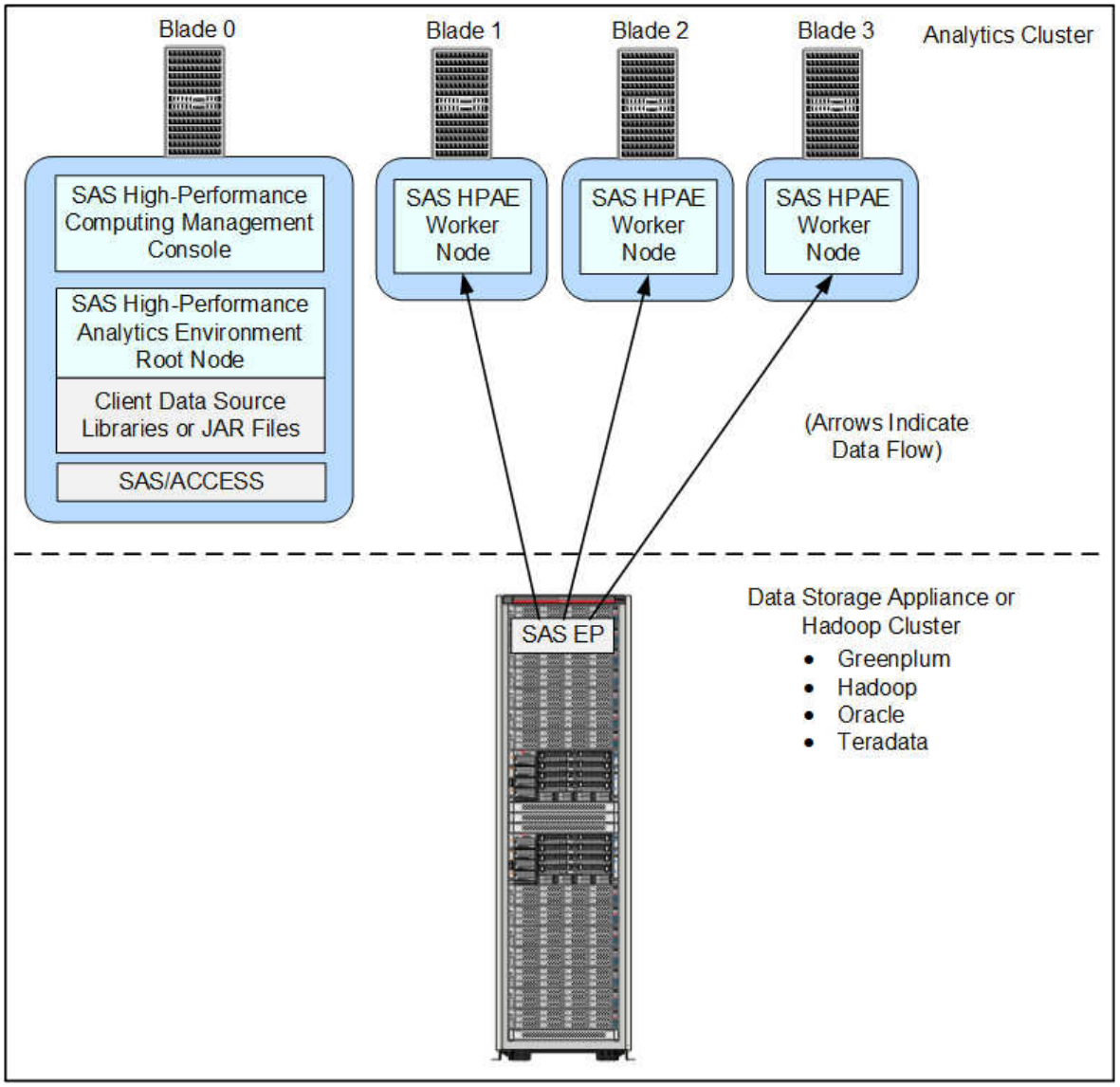


The serial connection between the analytics cluster and your data store is achieved by using the SAS/ACCESS Interface. SAS/ACCESS is orderable in a deployment package that is specific for your data source. For more information, refer to the *SAS/ACCESS for Relational Databases: Reference*, available at **http://support.sas.com/documentation/onlinedoc/access/index.html**.

Analytics Cluster Remote from Your Data Store (Parallel Connection)

The following figure shows the analytics cluster using a parallel connection to your remote data store:

Figure 1.4 *Analytics Cluster Remote from Your Data Store (Parallel Connection)*



Together the SAS/ACCESS Interface and SAS Embedded Process provide a high-speed parallel connection that delivers data from your data source to the SAS-High Performance Analytics environment on the analytic cluster. These components are contained in a deployment package that is specific for your data source. For more information, refer to the *SAS In-Database Products: Administrator's Guide*, available at `http://support.sas.com/documentation/onlinedoc/indbtech/index.html`.

Deploying the Infrastructure

Overview of Deploying the Infrastructure

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

1. Create a SAS Software Depot.
2. Check for documentation updates.
3. Prepare your analytics cluster.
4. (Optional) Deploy SAS High-Performance Computing Management Console.
5. (Optional) Deploy Hadoop.
6. Configure your data storage.
7. Deploy the SAS High-Performance Analytics environment.

The following sections provide a brief description of each of these tasks. Subsequent chapters in the guide provide the step-by-step instructions.

Step 1: 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 2: Check for Documentation Updates

It is very important to check for late-breaking installation information in SAS Notes and also to review the system requirements for your SAS software.

- SAS Notes

  Go to this web page and click **Outstanding Alert Status Installation Problems**:

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

- system requirements

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

Step 3: Prepare Your Analytics Cluster

Preparing your analytics cluster 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 is required 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 15.

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.

For example, 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 27.

Step 5: (Optional) Deploy Hadoop

If your site wants to use Hadoop as the co-located data store, then you can install and configure SAS High-Performance Deployment of Hadoop or use one of the supported Hadoop implementations.

For more information, see Chapter 4, “Deploying Hadoop,” on page 43.

Step 6: Configure Your Data Provider

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 cluster or data appliance.

For more information, see Chapter 5, “Configuring Your Data Provider,” on page 61.

Step 7: 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.

Software for the root node is deployed on the first host. Software for a worker node is installed on each remaining machine in the cluster or database appliance.

For more information, see Chapter 6, “Deploying the SAS High-Performance Analytics Environment,” on page 73.

2

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

Infrastructure Deployment Process Overview

Preparing your analytics cluster is the third of seven steps required to install and configure the SAS High-Performance Analytics infrastructure.

1. Create a SAS Software Depot.
2. Check for documentation updates.
3. **Prepare your analytics cluster.**
4. (Optional) Deploy SAS High-Performance Computing Management Console.
5. (Optional) Deploy Hadoop.
6. Configure your data provider.
7. Deploy the SAS High-Performance Analytics environment.

System Settings for the 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 27 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 19.

- Hadoop user account

  For more information about Hadoop user accounts, see "Preparing to Deploy Hadoop" on page 20.

- 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 23.

Preparing to Install SAS High-Performance Computing Management Console

User Account Considerations for the Management Console

SAS High-Performance Computing Management Console is installed from either an RPM or a tarball 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 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 17.
- 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 Deploy Hadoop

User Accounts for 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.

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 38.

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.

Install a Java Runtime Environment

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. 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/TKGrid/bin/simsh mkdir /opt/java
/opt/TKGrid/bin/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 HadoopDirectory 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.

Default Directory Location	Description
`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 50.

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 the Management Console" on page 19.

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.

Additional Prerequisite for Greenplum Deployments

For deployments that rely on Greenplum data appliances, the SAS High-Performance Analytics environment requires that you also deply the appropriate SAS/ACCESS interface and SAS Embedded Process that SAS supplies with SAS In-Database products. For more information, see *SAS In-Database Products: Administrator's Guide*, available at **http://support.sas.com/documentation/onlinedoc/indbtech/index.html**.

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	
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	

3

Deploying SAS High-Performance Computing Management Console

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. Create a SAS Software Depot.

2. Check for documentation updates.

3. Prepare your analytics cluster.

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

5. (Optional) Deploy Hadoop.

6. Configure your data provider.

7. Deploy the SAS High-Performance Analytics environment.

Benefits of the 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 38.

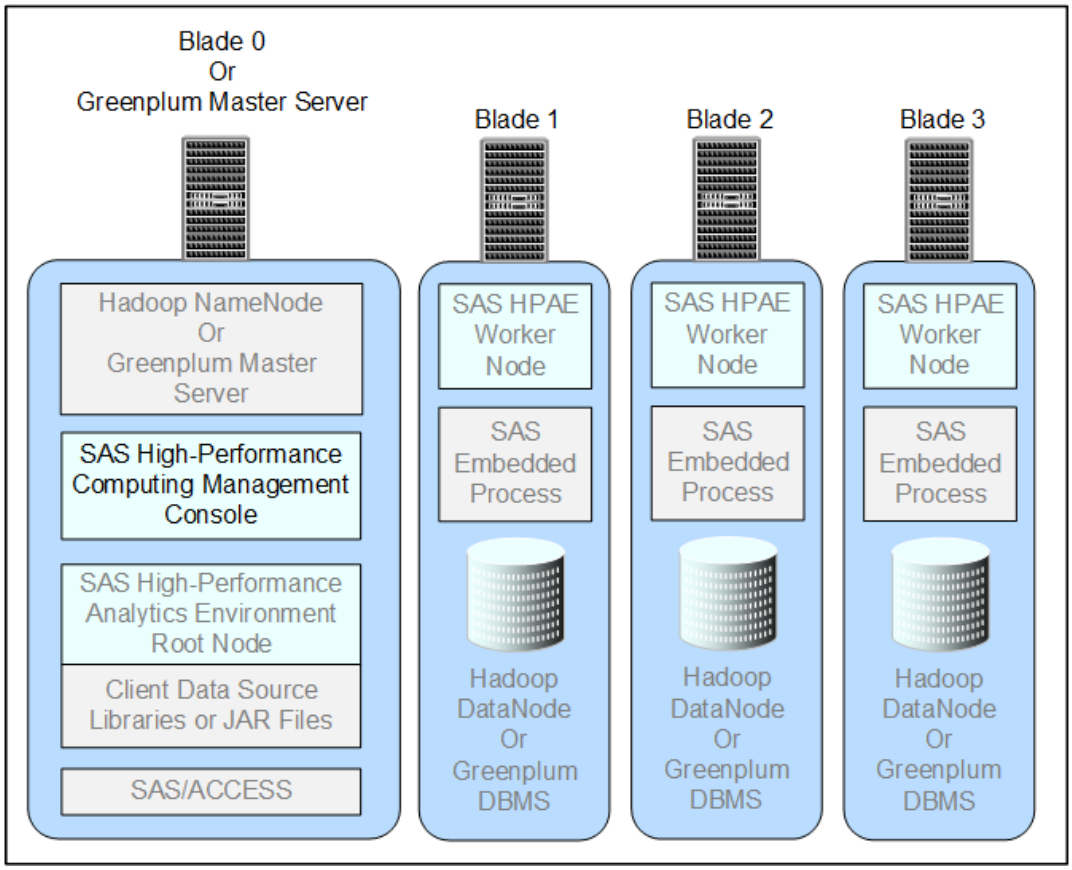
SAS High-Performance Computing Management Console enables you to perform these tasks from one location. When you create *new* user accounts using SAS High-Performance Computing Management Console, the console propagates the public key across all the machines in the cluster in a single operation.

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 the 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*



Installing the Management Console

There are two ways to install SAS High-Performance Computing Management Console.

Install SAS High-Performance Computing Management Console Using RPM

To install SAS High-Performance Computing Management Console using RPM, follow these steps:

Note: For information about updating the console, see "Updating the SAS High-Performance Analytics Infrastructure" on page 99.

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 19.
2. Log on to the target machine as root.
3. In your SAS Software Depot, locate the **`standalone_installs/SAS_High-Performance_Computing_Management_Console/2_5/Linux_for_x64`** directory.
4. Enter one of the following commands:
   - To install in the default location of **`/opt`**:

     ```
     rpm -ivh sashpcmc*
     ```

   - To install in a location of your choice:

     ```
     rpm -ivh --prefix=directory sashpcmc*
     ```

     where *directory* is an absolute path where you want to install the console.
5. Proceed to the topic "Configure the Management Console" on page 31.

Install the Management Console Using tar

Some versions of Linux use different RPM libraries and require and alternative means to Install SAS High-Performance Computing Management Console. Follow these steps to install the management console using tar:

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 19.

2 Log on to the target machine as root.

3 In your SAS Software Depot, locate the **standalone_installs/SAS_High-Performance_Computing_Management_Console/2_5/Linux_for_x64** directory.

4 Copy sashpcmc-2.5.tar.gz to the location where you want to install the management console.

5 Run the following command:

```
tar -xzvf sashpcmc-2.5.tar.gz
```

tar extracts the contents into a directory called **sashpcmc**.

6 Proceed to the topic "Configure the Management Console" on page 31.

Configure the 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 32. 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** for the script to 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.

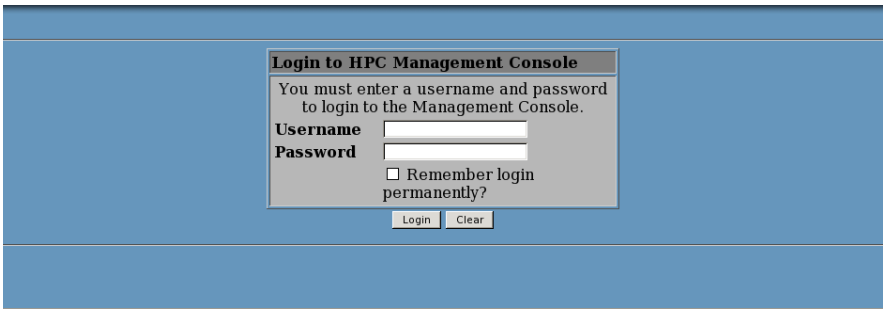
7 To start the SAS High-Performance Computing Management Console server, enter the following command from any directory:

```
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.

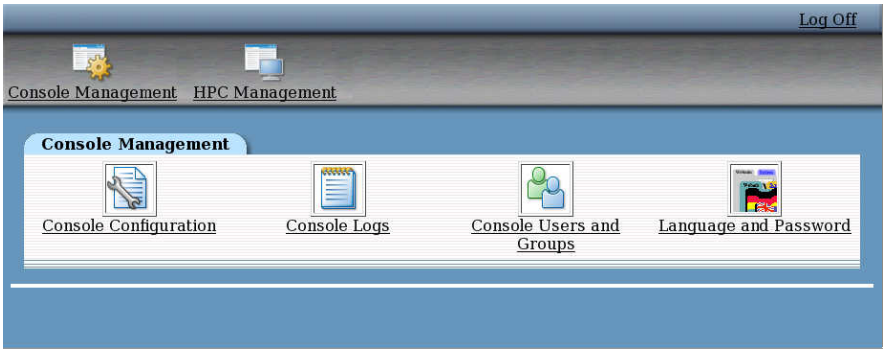
For example: **https://myserver.example.com:10020**

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:

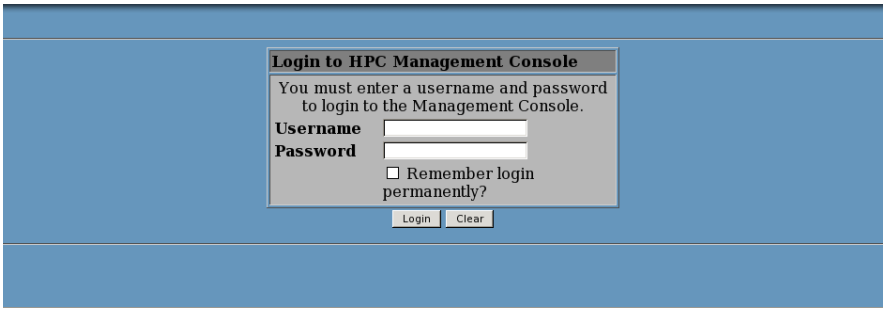
```
service sashpcmc status
```

(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 32.

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.

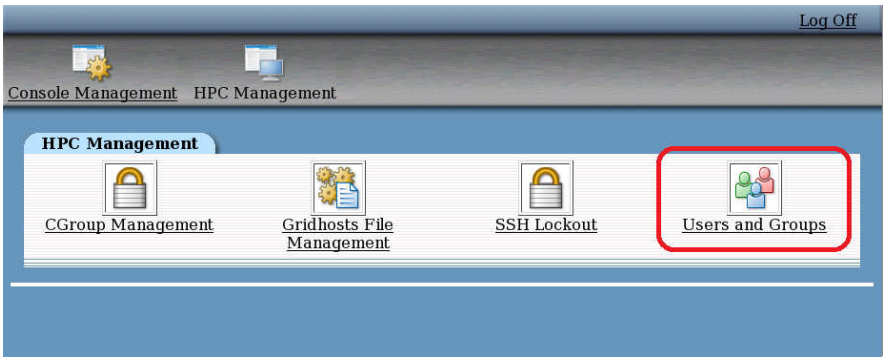


3 Log on to SAS High-Performance Computing Management Console.

The Console Management page appears.

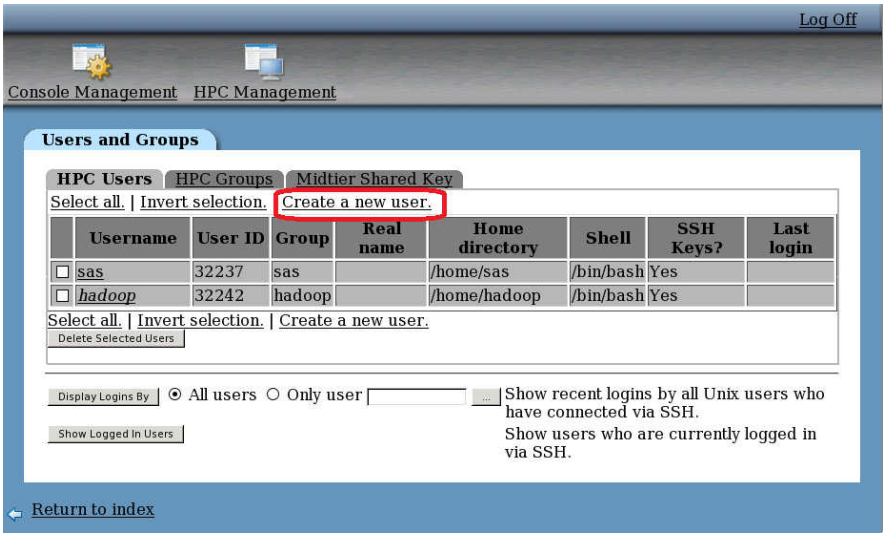
4 Click **HPC Management**.

The HPC Management page appears.



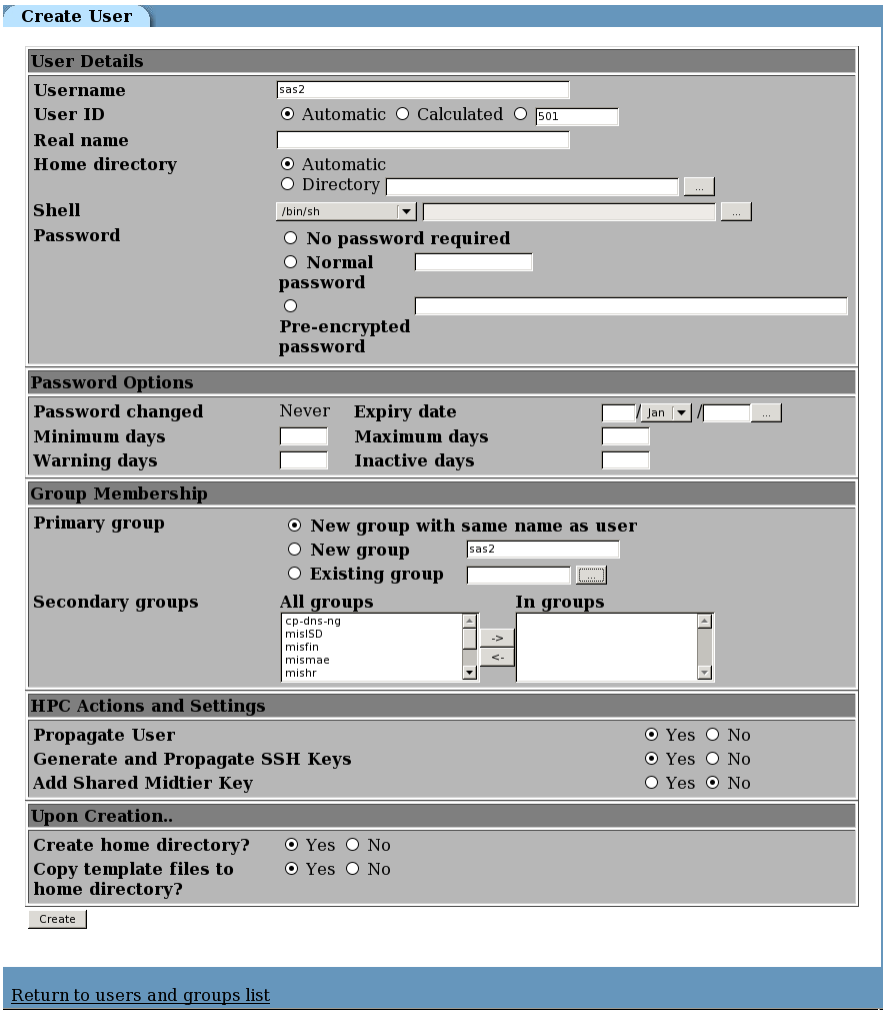
5 Click **Users and Groups**.

The Users and Groups page appears.



6 Click **Create a new user**.

The Create User page appears.



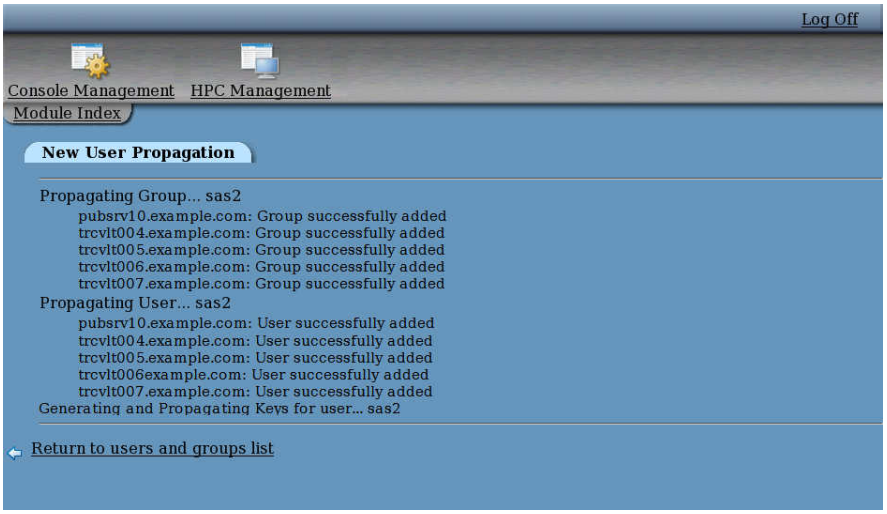
7 Enter information for the new user, using the security policies in place at your site.

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:

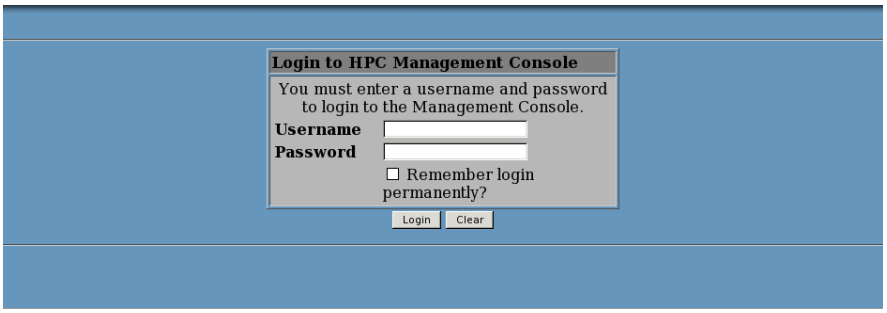
```
service sashpcmc status
```

For more information, see To start the SAS High-Performance Computing Management Console server on page 32.

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.

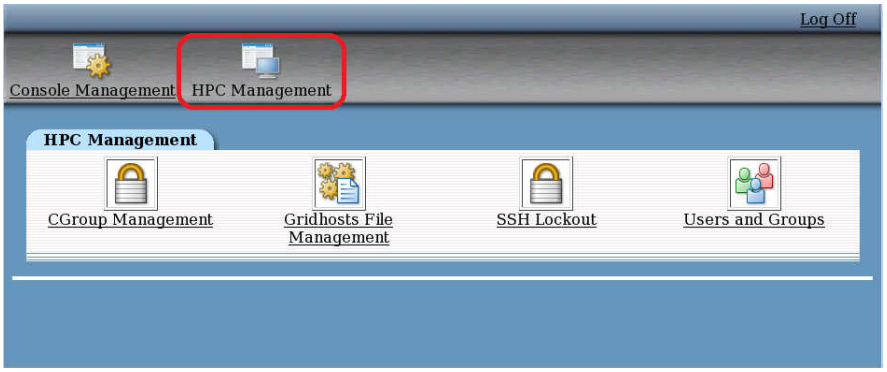
For example: `http://myserver.example.com:10020`

The Login page appears.



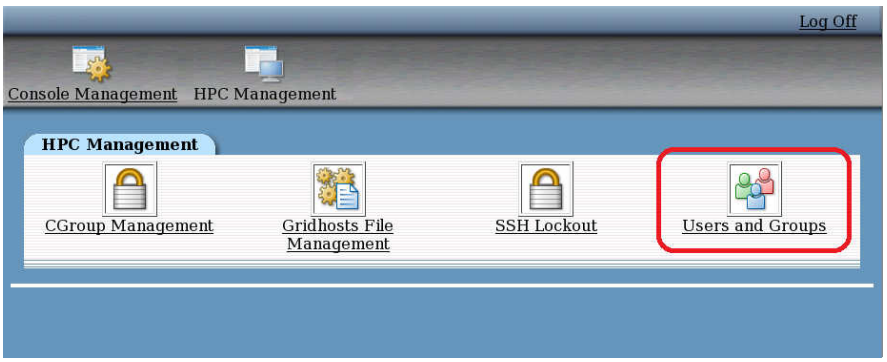
3 Log on to SAS High-Performance Computing Management Console.

The Console Management page appears.



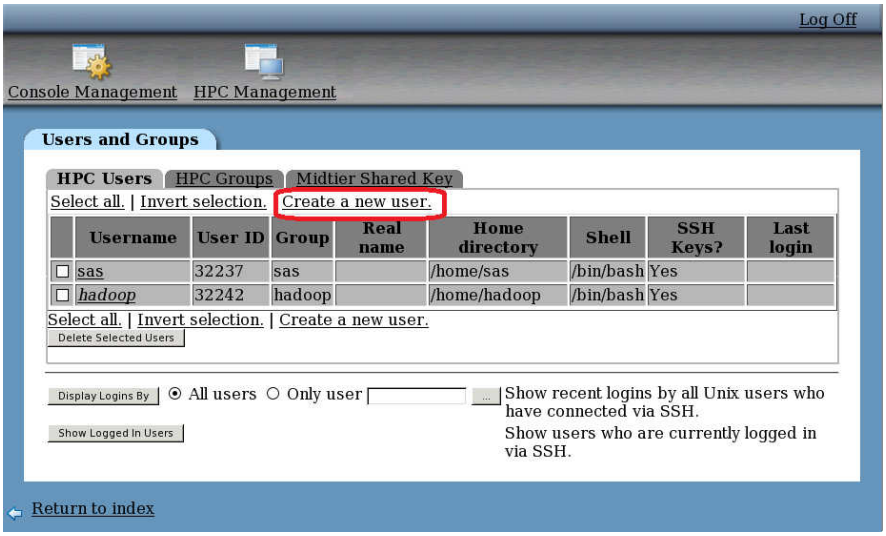
4 Click **HPC Management**.

The Console Management page appears.



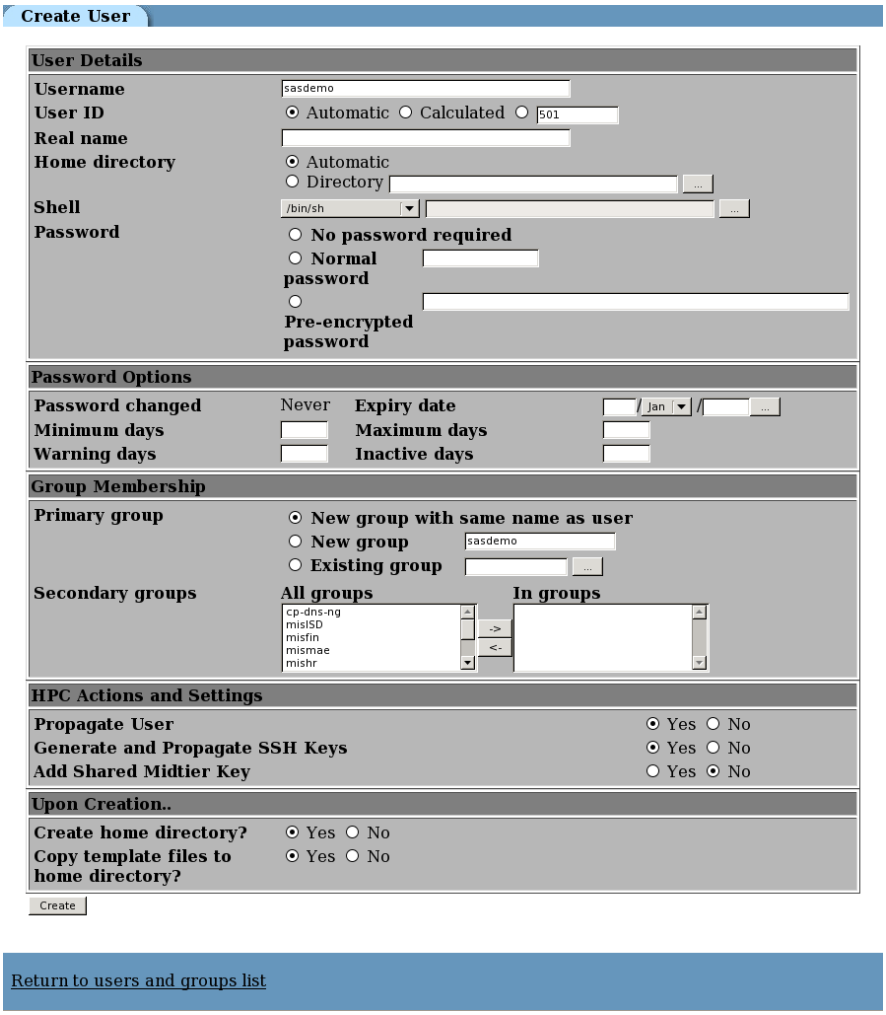
5 Click **Users and Groups**.

The Users and Groups page appears.



6 Click **Create a new user**.

The Create User page appears.



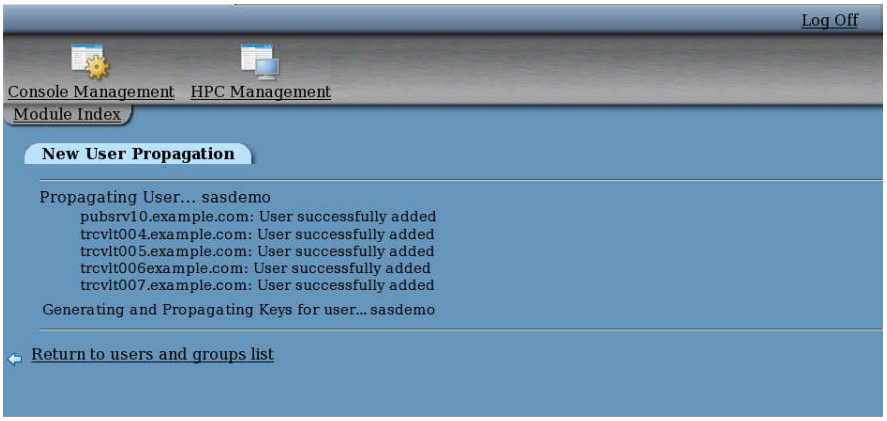
7 Enter information for the new user, using the security policies in place at your site.

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.


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

4

Deploying Hadoop

Infrastructure Deployment Process Overview

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

1. Create a SAS Software Depot.
2. Check for documentation updates.
3. Prepare your analytics cluster.
4. (Optional) Deploy SAS High-Performance Computing Management Console.
5. **(Optional) Deploy Hadoop.**
6. Configure your data provider.
7. Deploy the SAS High-Performance Analytics environment.

Overview of Deploying Hadoop

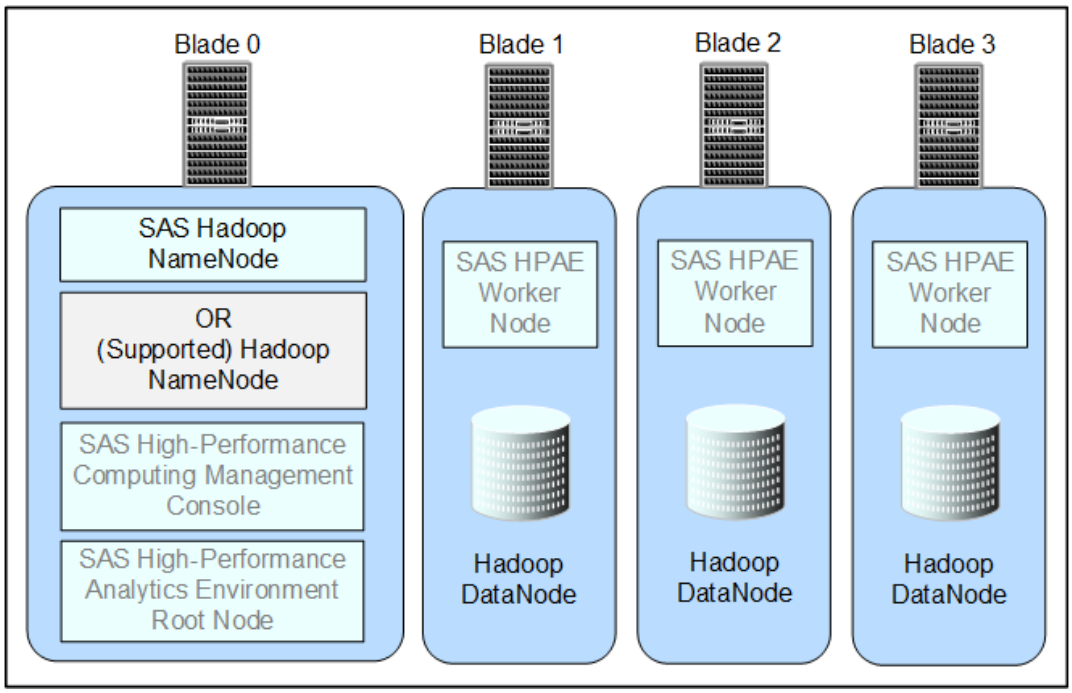
The SAS High-Performance Analytics environment relies on a massively parallel distributed database management system (Greenplum) or a Hadoop Distributed File System.

If you choose to use Hadoop, you have the option of using a Hadoop supplied by SAS, or using another supported Hadoop:

- “Deploying SAS High-Performance Deployment of Hadoop” on page 45.
- “Configuring Existing Hadoop Clusters” on page 53.

Deploying 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 *Analytics Cluster Co-Located on the Hadoop Cluster*



Deploying SAS High-Performance Deployment of Hadoop

Overview of Deploying SAS High-Performance Deployment of Hadoop

The steps required to deploy the SAS High-Performance Deployment of Hadoop consist of:

- “Install SAS High-Performance Deployment of Hadoop” on page 46.

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_5/Linux_for_x64/sashadoop.tar.gz`

1. Make sure that you have reviewed all of the information contained in the section "Preparing to Deploy Hadoop" on page 20.
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 Hadoop" on page 20.
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 46 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 25.

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 50.
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 If your deployment includes SAS/ACCESS Interface to Hadoop, install the appropriate SAS Embedded Process on your Hadoop machine cluster. For more information, see *SAS In-Database Products: Administrator's Guide*, available at **`http://support.sas.com/documentation/onlinedoc/indbtech/index.html`**.

10 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 50.
- If this is a new deployment of Hadoop, see "Format the Hadoop NameNode" on page 51.

(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.

For example:

```
<property>
  <name>dfs.data.dir</name>
  <value>/local/hadoop/hadoop-data,/data/dn</value>
</property>
```

4 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 103.

5 Restart Hadoop with the following command:

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

6 Proceed to "Format the Hadoop NameNode" on page 51.

Format the Hadoop NameNode

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

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

```
su - hadoop
```

2 Export the HADOOP_HOME environment variable.

For example:

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

3 Format the NameNode:

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

4 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
************************************************************/
```

5 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.

6 Create two directories in HDFS that permit Read and Write access for all users:

```
/hadoop-install-dir/hadoop-0.23.1/bin/hadoop fs -mkdir /vapublic
/hadoop-install-dir/hadoop-0.23.1/bin/hadoop fs -mkdir /hps
/hadoop-install-dir/hadoop-0.23.1/bin/hadoop fs -chmod 1777 /vapublic
/hadoop-install-dir/hadoop-0.23.1/bin/hadoop fs -chmod 777 /hps
```

7 Proceed to “Validate Your Hadoop Deployment” on page 52.

Validate Your Hadoop Deployment

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.

Configuring Existing Hadoop Clusters

Overview of Configuring Existing Hadoop Clusters

If your site uses a Hadoop implementation that is supported, then you can configure your Hadoop cluster for use with the SAS High-Performance Analytics environment.

The steps needed to configure your existing Hadoop cluster consist of:

1. Make sure that your Hadoop deployment meets the analytic environment prerequisites. For more information, see "Prerequisites for Existing Hadoop Clusters" on page 53
2. Follow steps specific to your implementation of Hadoop:
   - "Configuring the Existing Cloudera Hadoop Cluster" on page 54
   - "Configuring the Existing Hortonworks Data Platform Hadoop Cluster" on page 57
   - "Configuring the Existing Pivotal HD Hadoop Cluster" on page 59

Prerequisites for Existing Hadoop Clusters

The following is required for existing Hadoop clusters that will be configured for use with the SAS High-Performance Analytics environment:

- Each machine machine in the cluster must be able to resolve the host name of all the other machines.
- The NameNode and secondary NameNode are not defined as the same host.

- The NameNode host does not also have a DataNode configured on it.
- For Kerberos, in the SAS High-Performance Analytics environment, **`/etc/hosts`** must contain the machine names in the cluster in this order: short name, fully qualified domain name.
- Time must be synchronized across all machines in the cluster.
- If you are using SAS 9.4 (TS1M0), the user running HDFS services in the analytics environment requires sudo privileges for a few commands. (This does **not** apply to the first maintenance release of SAS 9.4 and later.)

Configuring the Existing Cloudera Hadoop Cluster

Managing Cloudera Configuration Priorities

Cloudera uses the Linux `alternatives` command for client configuration files. Therefore, make sure that the client configuration path has the highest priority for all machines in the cluster. (Often, the mapreduce client configuration has a higher priority over the hdfs configuration.)

If the output of the command `alternatives –display hadoop-conf` returns the Cloudera server configuration or mapreduce client configuration has priority over the client configuration, you will experience problems because SAS makes additions to the client configuration. For more information about `alternatives`, refer to its man page.

Configure the Existing Cloudera Hadoop Cluster

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

1 Untar the SAS High-Performance Deployment for Hadoop tarball, and propagate three files (identified below) on every machine in your Cloudera Hadoop cluster:

   a Navigate to the SAS High-Performance Deployment for Hadoop tarball in your SAS Software depot:

   ```
   cd depot-installation-location/standalone_installs/
   SAS_High_Performance_Hadoop_Deployment/2_5/Linux_for_x64/
   ```

b Untar sashadoop.tar.gz:

```
tar xzf sashadoop.tar.gz
```

c Locate sas.lasr.jar and sas.hadoop.lasr.jar and propagate these two JAR files to every machine in the Cloudera Hadoop cluster into the CDH library path.

TIP You can issue a single **simcp** command to propagate JAR files across all machines in the cluster. For example:

```
/opt/TKGrid/bin/simcp sas.lasr.jar
/opt/cloudera/parcels/CDH-4.4.0-1.cdh4.4.0.p0.39/lib/hadoop/lib/
/opt/TKGrid/bin/simcp sas.lasr.hadoop.jar
/opt/cloudera/parcels/CDH-4.4.0-1.cdh4.4.0.p0.39/lib/hadoop/lib/
```

For more information, see Appendix 3, “SAS High-Performance Analytics Infrastructure Command Reference,” on page 103.

d Locate saslasrfd and propagate this file to every machine in the Cloudera Hadoop cluster into the CDH **bin** directory. For example:

```
/opt/TKGrid/bin/simcp saslasrfd
/opt/cloudera/parcels/CDH-4.4.0-1.cdh4.4.0.p0.39/lib/hadoop/bin/
```

2 Log on to the Cloudera Manager as an administrator.

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

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

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

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

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

```
<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>
<property>
<name> dfs.namenode.fs-limits.min-block-size</name>
<value>0</value>
</property>
```

6 Restart all Cloudera Manager services.

7 Create and set the mode for the **/test** directory in HDFS for testing. You might need to set HADOOP_HOME first, and you must run the following commands as the user running HDFS (normally, the hdfs user).

8 If needed, set the following environment variables before running the Hadoop commands.

```
export HADOOP_HOME=/opt/cloudera/parcels/CDH-4.4.0-1.cdh4.4.0.p0.39/lib/hadoop
```

9 Run the following commands to create the /test directory in HDFS. This directory is to be used for testing the cluster with SAS test jobs.

```
$HADOOP_HOME/bin/hadoop fs -mkdir /test
```

```
$HADOOP_HOME/bin/hadoop fs -chmod 777 /test
```

10 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>
```

11 Add the location of JAVA_HOME to the Client Environment Safety Valve for hadoop-env.sh. For example:

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

12 Save your changes and deploy the client configuration to each host in the cluster.

13 Make sure that the client configuration path has the highest priority for all machines in the cluster. For more information, see "Managing Cloudera Configuration Priorities" on page 54.

Configuring the Existing Hortonworks Data Platform Hadoop Cluster

Use the Ambari interface to configure your existing Hortonworks Data Platform deployment to interoperate with the SAS High-Performance Analytics environment.

1 Log on to Ambari as an administrator, and stop all HDP services.

2 Untar the SAS High-Performance Deployment for Hadoop tarball, and propagate three files (identified below) on every machine in your Cloudera Hadoop cluster:

a Navigate to the SAS High-Performance Deployment for Hadoop tarball in your SAS Software depot:

```
cd depot-installation-location/standalone_installs/
SAS_High_Performance_Hadoop_Deployment/2_5/Linux_for_x64/
```

b Untar sashadoop.tar.gz:

```
tar xzf sashadoop.tar.gz
```

c Locate sas.lasr.jar and sas.hadoop.lasr.jar and propagate these two JAR files to every machine in the HDP cluster into the HDP library path.

> **TIP** You can issue a single **simcp** command to propagate JAR files across all machines in the cluster. For example:
>
> ```
> /opt/TKGrid/bin/simcp sas.lasr.jar /usr/lib/hadoop/lib/
> /opt/TKGrid/bin/simcp sas.lasr.hadoop.jar /usr/lib/hadoop/lib/
> ```

For more information, see Appendix 3, “SAS High-Performance Analytics Infrastructure Command Reference,” on page 103.

d Locate saslasrfd and propagate this file to every machine in the HDP cluster into the HDP **bin** directory. For example:

```
/opt/TKGrid/bin/simcp saslasrfd /usr/lib/hadoop/bin/
```

3 In the Ambari interface, create a custom hdfs-site.xml and add the following properties:

dfs.namenode.plugins
com.sas.lasr.hadoop.NameNodeService

dfs.datanode.plugins
com.sas.lasr.hadoop.DataNodeService

com.sas.lasr.hadoop.fileinfo
ls -l {0}

com.sas.lasr.service.allow.put
true

com.sas.lasr.hadoop.service.namenode.port
15452

com.sas.lasr.hadoop.service.datanode.port
15453

dfs.namenode.fs-limits.minblock.size
0

4 Save the properties and start the HDFS service.

5 Run the following commands as the hdfs user to create the **/test** directory in HDFS. This directory is used for testing your cluster with SAS test jobs.

```
hadoop fs -mkdir /test
hadoop fs -chmod 777 /test
```

Configuring the Existing Pivotal HD Hadoop Cluster

Use the Pivotal Command Center (PCC) to configure your existing Pivotal HD deployment to interoperate with the SAS High-Performance Analytics environment.

1 Log on to PCC as gpadmin. (The default password is gpadmin.)

2 Untar the SAS High-Performance Deployment for Hadoop tarball, and propagate three files (identified below) on every machine in your Cloudera Hadoop cluster:

   a Navigate to the SAS High-Performance Deployment for Hadoop tarball in your SAS Software depot:

   ```
   cd depot-installation-location/standalone_installs/
   SAS_High_Performance_Hadoop_Deployment/2_5/Linux_for_x64/
   ```

   b Untar sashadoop.tar.gz:

   ```
   tar xzf sashadoop.tar.gz
   ```

   c Locate sas.lasr.jar and sas.hadoop.lasr.jar and propagate these two JAR files to every machine in the Pivotal HD cluster into the library path.

   TIP You can issue a single **simcp** command to propagate JAR files across all machines in the cluster. For example:

   ```
   /opt/TKGrid/bin/simcp sas.lasr.jar /usr/lib/gphd/hadoop/lib/
   /opt/TKGrid/bin/simcp sas.lasr.hadoop.jar /usr/lib/gphd/hadoop/lib/
   ```

   For more information, see Appendix 3, “SAS High-Performance Analytics Infrastructure Command Reference,” on page 103.

   d Locate saslasrfd and propagate this file to every machine in the Pivotal HD cluster into the Pivotal HD**bin** directory. For example:

   ```
   /opt/TKGrid/bin/simcp saslasrfd /usr/lib/gphd/hadoop/bin/
   ```

3 In the PCC, for Yarn, make sure that Resource Manager, History Server, and Node Managers have unique host names.

4 In the PCC, make sure that the Zookeeper Server contains a unique host name.

5 Add the following properties for SAS for the HDFS configuration in the file, hdfs-site.xml:

```
<property>
<name>dfs.datanode.plugins</name>
<value>com.sas.lasr.hadoop.DataNodeService</value>
</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>
<property>
<name> dfs.namenode.fs-limits.min-block-size</name>
<value>0</value>
</property>
```

6 Save your changes and deploy.

7 Restart your cluster using PCC and verify that HDFS is running in the dashboard.

8 Run the following commands as the gpadmin user to create the **`/test`** directory in HDFS. This directory is used for testing your cluster with SAS test jobs.

```
hadoop fs -mkdir /test
```

```
hadoop fs -chmod 777 /test
```

5

Configuring Your Data Provider

Infrastructure Deployment Process Overview

Configuring your data storage is the sixth of seven steps for deploying the SAS High-Performance Analytics infrastructure.

1. Create a SAS Software Depot.
2. Check for documentation updates.
3. Prepare your analytics cluster.
4. (Optional) Deploy SAS High-Performance Computing Management Console.
5. (Optional) Deploy Hadoop.
6. **Configure your data provider.**
7. Deploy the SAS High-Performance Analytics environment.

Overview of Configuring Your Data Provider

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

The topics that follow describe how you configure the data sources that you are using with the analytics environment:

- "Recommended Database Names" on page 65
- "Preparing the Greenplum Database for SAS" on page 66
- "Overview of Preparing Your Data Provider for a Parallel Connection with SAS" on page 69

The figures that follow illustrate the various ways in which you can configure data access for the analytics environment:

- Analytics cluster co-located on the Hadoop cluster or Greenplum data appliance on page 63
- Analytics cluster remote from your data store (serial connection) on page 64
- Analytics cluster remote from your data store (parallel Connection) on page 65

Figure 5.1 *Analytics Cluster Co-Located on the Hadoop Cluster or Greenplum Data Appliance*

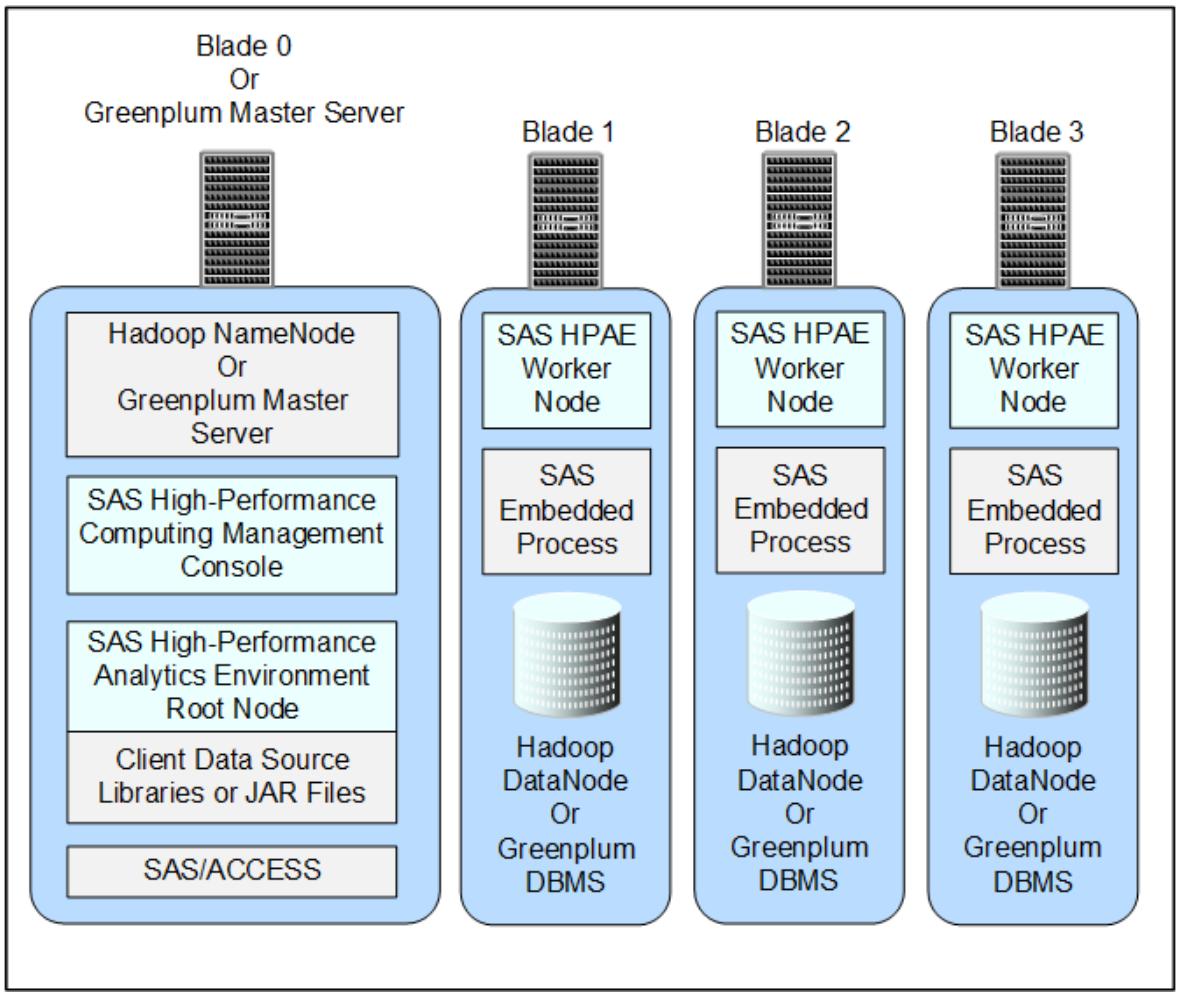


Figure 5.2 *Analytics Cluster Remote from Your Data Store (Serial Connection)*

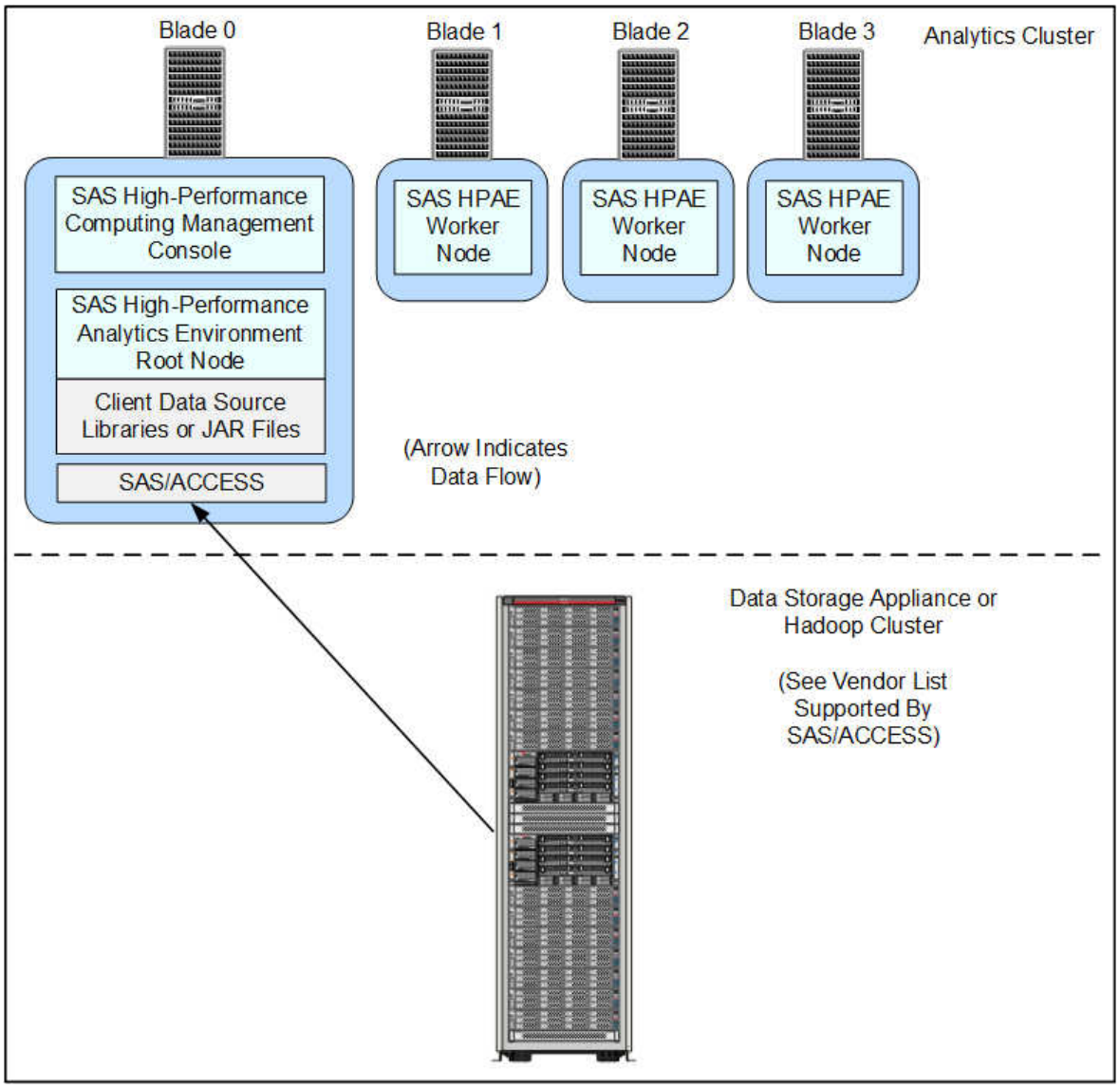
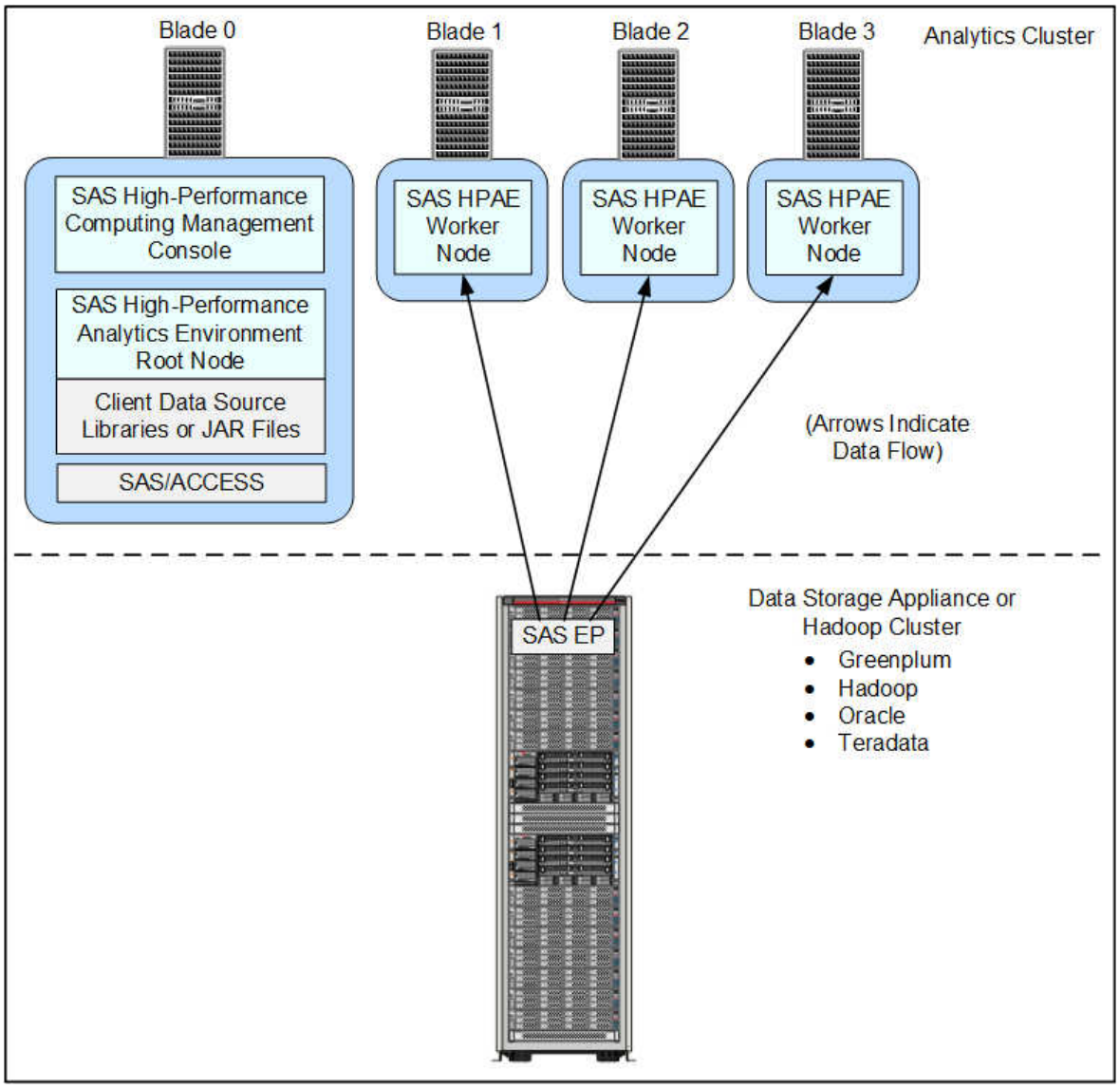


Figure 5.3 *Analytics Cluster Remote from Your Data Store (Parallel Connection)*



Recommended Database Names

SAS solutions, such as SAS Visual Analytics, that rely on a co-located data provider can make use of two database instances.

The first instance often already exists and is expected to have your operational or transactional data that you want to explore and analyze.

A second database instance is used to support the self-service data access features of SAS Visual Analytics. This database is commonly named “vapublic,” but you can specify a different name if you prefer. Keep these names handy, as the SAS Deployment Wizard prompts you for them when deploying your SAS solution.

Preparing the Greenplum Database for SAS

Overview of Preparing the Greenplum Database for SAS

The steps required to configure your Greenplum database for the SAS High-Performance Analytics environment consist of the following:

1 Associate users with a group role.

2 Configure the SAS/ACCESS Interface to Greenplum.

3 Install the SAS Embedded Process for Greenplum.

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 role group 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 rights on the schema where the SAS Embedded Process has been published. For more information, see *SAS In-Database Products: Administrator's Guide*, available at **http://support.sas.com/documentation/onlinedoc/indbtech/index.html**.

For example:

```
GRANT ALL ON SCHEMA SASLIB TO sas_cust_group;
```

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 Set the ODBCHOME environment variable to your ODBC home directory.

2 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.

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

4 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

Install the SAS Embedded Process for Greenplum

If you have not done so already, install the appropriate SAS Embedded Process on your Greenplum data appliance. For more information, see *SAS In-Database Products: Administrator's Guide*, available at http://support.sas.com/documentation/onlinedoc/indbtech/index.html.

Note: While following the instructions in the *SAS In-Database Products: Administrator's Guide*, there is no need to run the %INDGP_PUBLISH_COMPILEUDF macro. All the other steps, including running the %INDGP_PUBLISH_COMPILEUDF_EP macro, are required.

Preparing Your Data Provider for a Parallel Connection with SAS

Overview of Preparing Your Data Provider for a Parallel Connection with SAS

Before you can configure the SAS High-Performance Analytics environment to use the SAS Embedded Process for a parallel connection with your data store, you must locate particular JAR files and gather particular information about your data provider. If you are using a Hadoop not supplied by SAS, then you must also complete a few configuration steps.

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

1 “Preparing Your Data Provider for a Parallel Connection with SAS” on page 69.

2 “Prepare for a Greenplum Data Appliance” on page 70.

3 “Prepare for an Oracle Exadata Appliance” on page 71.

4 “Prepare for a Teradata Managed Server Cabinet” on page 72.

Prepare for Hadoop

Before you can configure the SAS High-Performance Analytics environment to use the SAS Embedded Process for a parallel connection with your Hadoop data store, there are certain requirements that must be met.

1 Depending on the Hadoop vendor, copy the following JAR files into a directory on blade 0 in the analytics cluster. Record the directory where you copy these JAR files in the table in step 2:

   - For the list of Cloudera common and core JAR files, refer to “Cloudera Hadoop JAR Files” on page 92.

- For the list of Hortonworks Data Platform common and core JAR files, refer to "Hortonworks Data Platform Hadoop" on page 94.

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

Table 5.1 *Record the Location of the Hadoop JAR Files Required by SAS*

Example	Actual Path of the Required Hadoop JAR Files on Your System
/opt/hadoop_jars (common and core JAR files)	
/opt/hadoop_jars/MR1 (Map Reduce JAR files)	
/opt/hadoop_jars/MR2 (Map Reduce JAR files)	

3 Record the path to the 64-bit Java Runtime Engine (JRE) required by SAS 9.4 in the table that follows:

Table 5.2 *Record the Location of the JRE*

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

Note: To determine the JRE version required by SAS and where to download it from, refer to `http://support.sas.com/resources/thirdpartysupport/v94/jres.html`.

Prepare for a Greenplum Data Appliance

Before you can configure the SAS High-Performance Analytics environment to use the SAS Embedded Process for a parallel connection with your Greenplum data appliance, there are certain requirements that must be met.

1 Install the Greenplum client on the Greenplum Master Server (blade 0) 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 5.3 *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 configure the SAS High-Performance Analytics environment to use the SAS Embedded Process for a parallel connection with your Oracle Exadata appliance, there are certain requirements that must be met.

1 Install the Oracle client on blade 0 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 5.4 *Record the Location of the Oracle Client*

Example	Actual Path of the Oracle Client on Your System
/usr/local/ora11gr2/product/11.2.0/cli 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 5.5 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 configure the SAS High-Performance Analytics environment to use the SAS Embedded Process for a parallel connection with your Teradata Managed Server Cabinet, there are certain requirements that must be met.

1 Install the Teradata client on blade 0 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 5.6 Record the Location of the Teradata Client

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

6

Deploying the SAS High-Performance Analytics Environment

Infrastructure Deployment Process Overview

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

1. Create a SAS Software Depot.

2. Check for documentation updates.

3. Prepare your analytics cluster.

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

5. (Optional) Deploy Hadoop.

6. Configure your data provider.

► **7. Deploy the SAS High-Performance Analytics environment.**

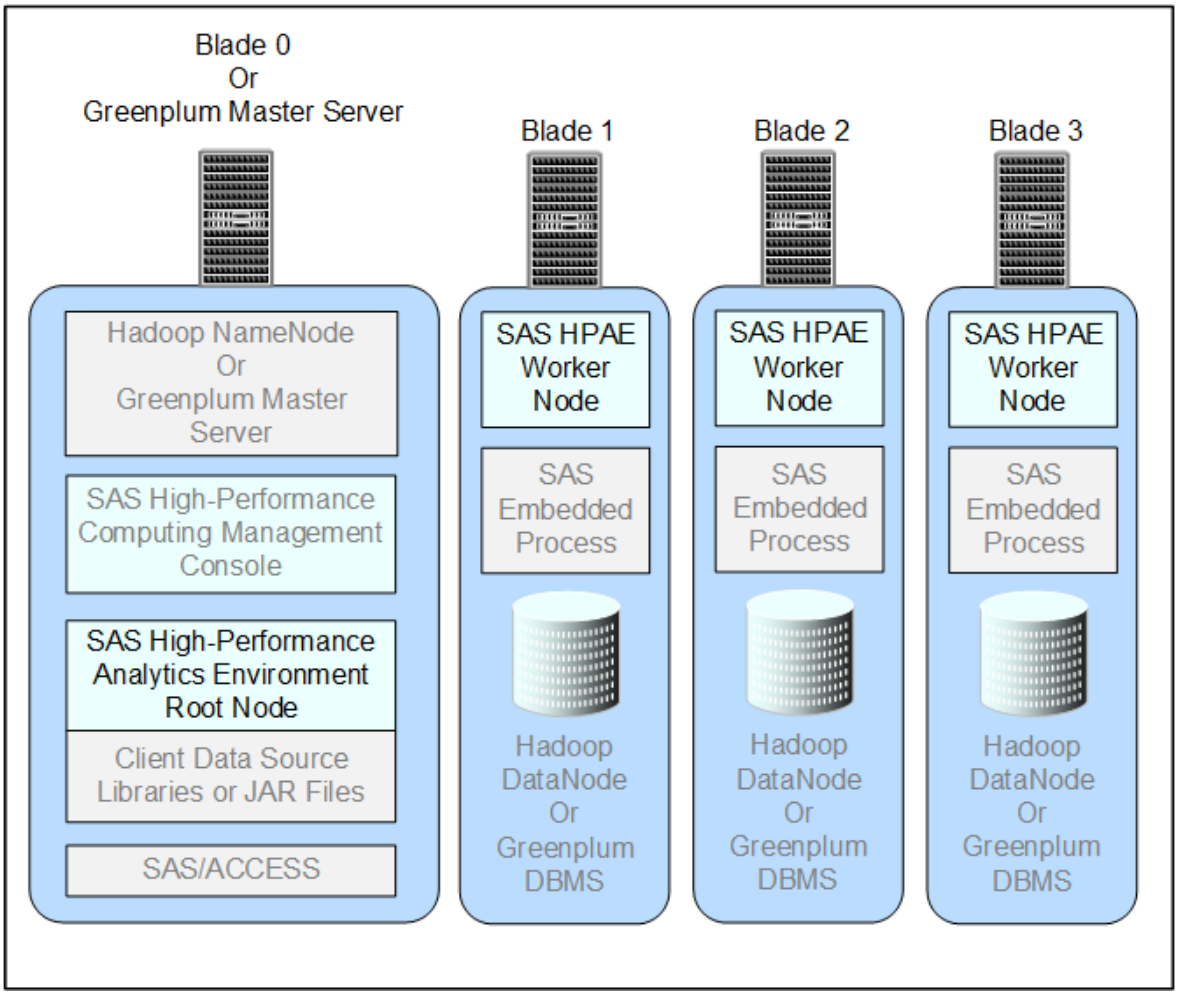
This chapter describes how to install and configure all of the components for the SAS High-Performance Analytics environment on the machines in the cluster.

Overview of Deploying the 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) or on the Master Server (Greenplum).

The following figure shows the SAS High-Performance Analytics environment co-located on your Hadoop cluster or Greenplum data appliance:

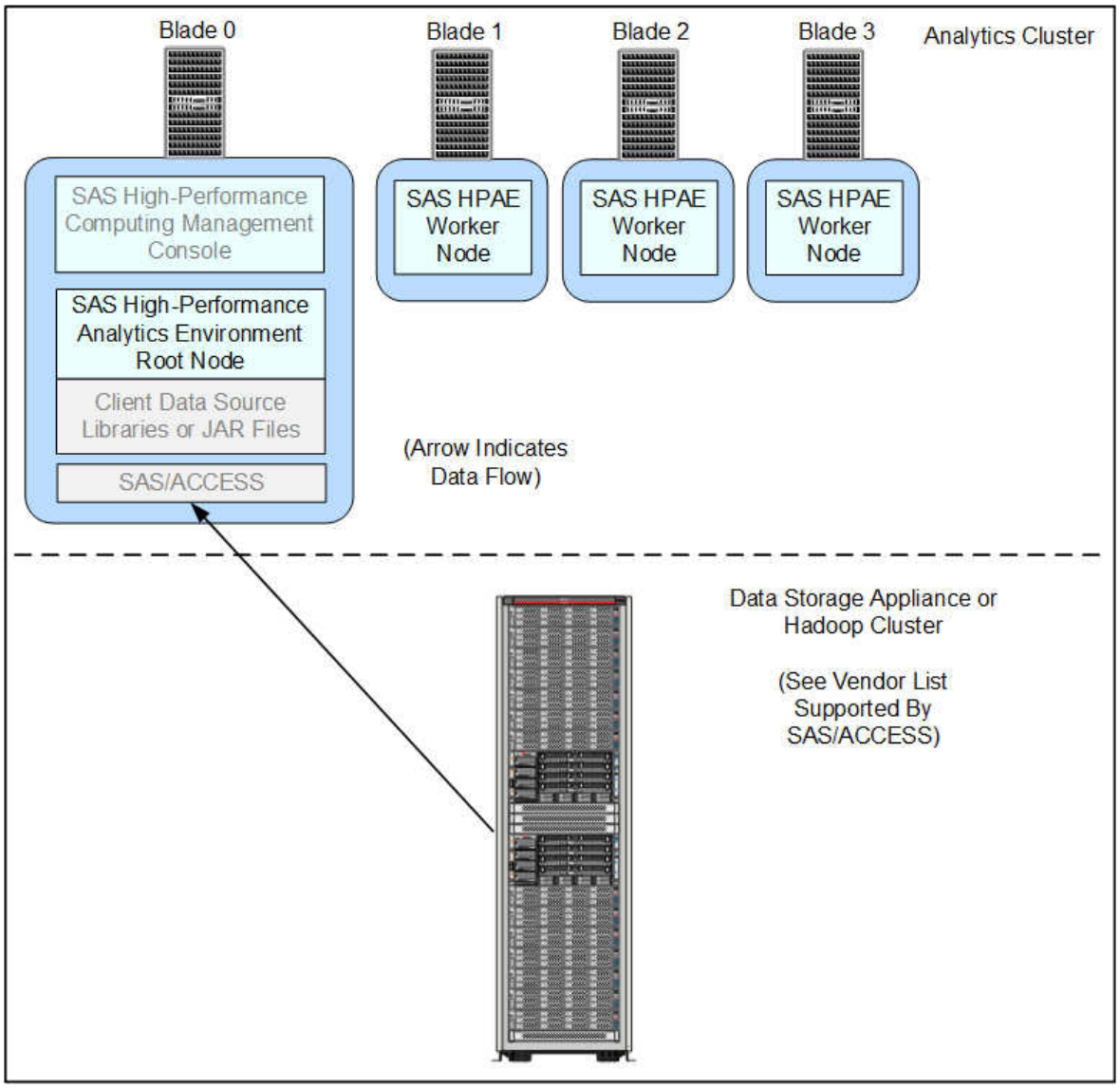
Figure 6.1 *Analytics Environment Co-Located on the Hadoop Cluster or Greenplum Data Appliance*



Note: For deployments that use Hadoop for the co-located data provider and access SASHDAT tables exclusively, SAS/ACCESS and SAS Embedded Process are not needed.

The following figure shows the SAS High-Performance Analytics environment using a serial connection through the SAS/ACCESS Interface to your remote data store:

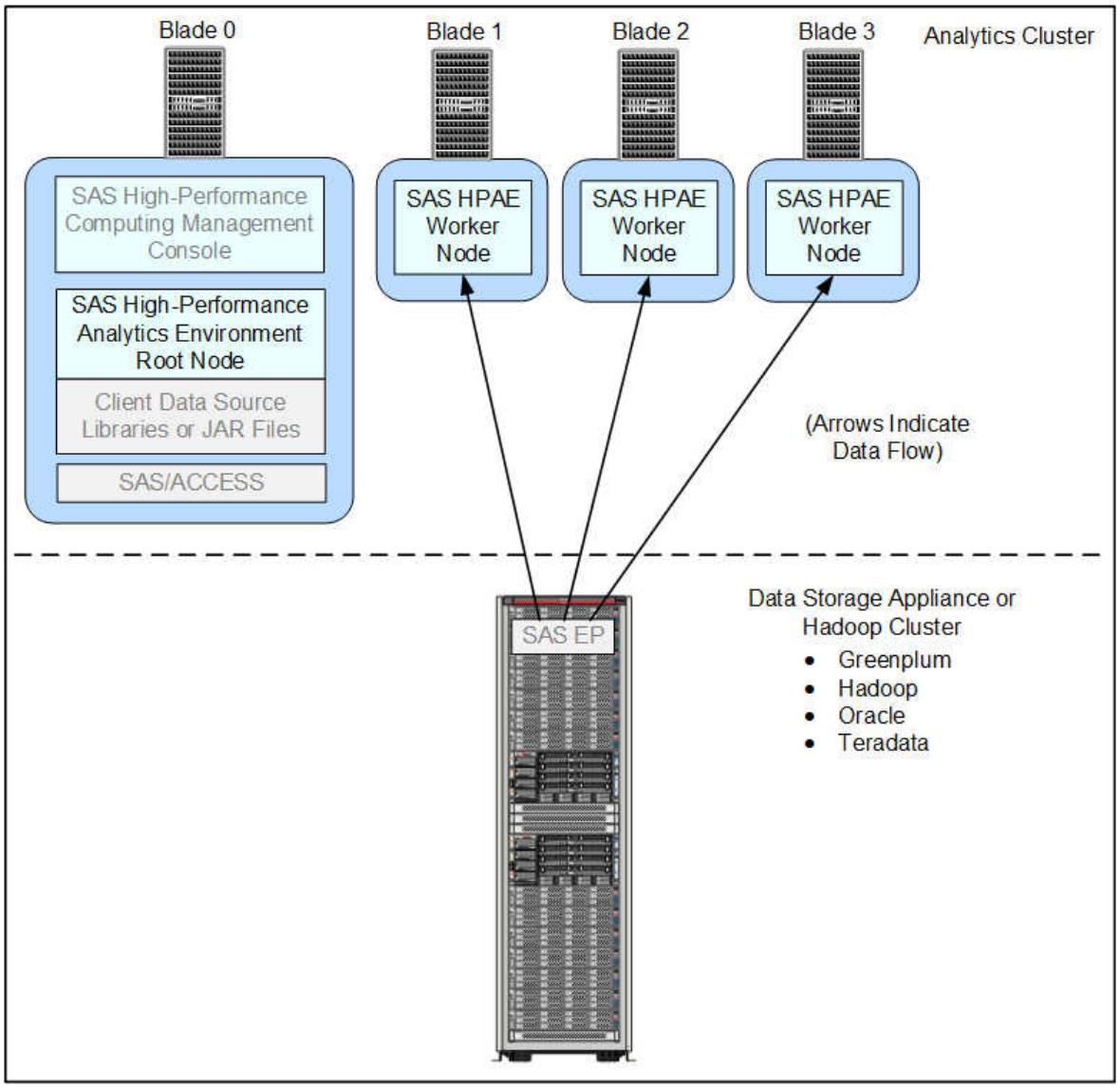
Figure 6.2 *Analytics Environment Remote from Your Data Store (Serial Connection)*



TIP There might be solution-specific criteria that you should consider when determining your analytics cluster location. For more information, see the installation or administration guide for your specific SAS solution.

The following figure shows the SAS High-Performance Analytics environment using a parallel connection through the SAS Embedded Process to your remote data store:

Figure 6.3 *Analytics Environment Remote from Your Data Store (Parallel Connection)*



Install the Analytics Environment

The SAS High-Performance Analytics environment components are installed with two shell scripts. Follow these steps to install:

1 Make sure that you have reviewed all of the information contained in the section "Preparing to Deploy Hadoop" on page 20.

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_5/Linux_for_x64`.

3 Copy the file that is appropriate for your operating system to the `/tmp` directory of the root node of the cluster:

   - Red Hat Linux (pre-version 6) and SUSE Linux 10:

     TKGrid_Linux_x86_64_rhel5.sh

   - Red Hat Linux 6 and other equivalent, kernel-level Linux systems:

     TKGrid_Linux_x86_64.sh

4 Copy TKTGDat.sh to the `/tmp` directory of the root node of the cluster.

   Note: TKTGDat.sh contains the SAS linguistic binary files required to perform text analysis in SAS LASR Analytic Server with SAS Visual Analytics and to run PROC HPTMINE and HPTMSCORE with SAS Text Miner.

5 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 23.

6 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.

7 Run the TKGrid shell script in this directory.

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

8 Respond to the prompts from the shell script:

Table 6.1 *Configuration Parameters for the TKGrid Shell Script*

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 remote process launcher command (Default is ssh).	If you want to use another program other than SSH to start processes on remote nodes, enter the program's absolute installation path.
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`**

Parameter	Description
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 (the value of the variable, HADOOP_HOME). If you are using SAS High-Performance Deployment of Hadoop, use the directory that you entered earlier in Step 3 on page 46. If your site does not use Hadoop, enter nothing and press the Enter or Return key.
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. The head node '**`head-node-machine-name`**' should be listed first.	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 runtime 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 24.

9 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.

10 Next, in the same directory from which you ran the TKGrid shell script, run TKTGDat.sh.

The shell script creates the **TKTGDat** subdirectory and places all files in that directory.

11 Respond to the prompts from the shell script:

Table 6.2 *Configuration Prompts for the TKTG Dat Shell Script*

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 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**).

12 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.

13 Proceed to “Validating the Analytics Environment Deployment” on page 87.

Configuring for a Remote Data Store

Overview of Configuring for a Remote Data Store

The process involved for configuring the SAS High-Performance Analytics environment for a remote data store consists of the following steps:

1 Prepare for the data provider that the analytics environment will query.

   For more information, see “Preparing Your Data Provider for a Parallel Connection with SAS” on page 69.

2 Review the considerations for configuring the analytics environment for use with a remote data store.

   For more information, see “How the Configuration Script Works” on page 82.

3 Configure the analytics environment for a remote data store.

   For more information, see “Configure for a Remote Data Store” on page 84.

How the Configuration Script Works

You configure the SAS High-Performance Analytics environment for a remote data store using a shell script. The script enables you to configure the environment for the various third-party data stores supported by the SAS Embedded Process.

The Analytics environment is designed on the principle, install once, configure many. For example, suppose that your site has three remote data stores from three different third-party vendors whose data you want to analyze. You run the analytics environment configuration script one time and provide the information for each data store vendor as

you are prompted for it. (When prompted for a data store vendor that you do not have, simply ignore that set of prompts.)

When you have different versions of the same vendor's data store, specifying the vendor's *latest* client data libraries usually works. However, this choice can be problematic for different versions of Hadoop, where a later set of JAR files is not typically backwardly compatible with earlier versions, or for sites that use Hadoop implementations from more than one vendor. (The configuration script does not delineate between different Hadoop vendors.) In these situations, you must run the analytics environment configuration script once for each different Hadoop version or vendor. As the configuration script creates a **`TKGrid_REP`** directory underneath the current directory, it is important to run the script a second time from a different directory.

To illustrate how you might manage configuring the analytics environment for two different Hadoop vendors, consider this example: suppose your site uses Cloudera Hadoop 4 and Hortonworks Data Platform 2. When running the analytics environment script to configure for Cloudera 4, you would create a directory similar to:

```
cdh4
```

When configuring the analytics environment for Cloudera, you would run the script from the **`cdh4`** directory. When complete, the script creates a **`TKGrid_REP`** child directory:

```
cdh4/TKGrid_REP
```

For Hortonworks, you would create a directory similar to:

```
hdp2
```

When configuring the analytics environment for Hortonworks, you would run the script from the **`hdp2`** directory. When complete, the script creates a **`TKGrid_REP`** child directory:

```
hdp2/TKGrid_REP
```

Configure for a Remote Data Store

To configure the High-Performance Analytics environment for a remote data store, follow these steps:

1 Make sure that you have reviewed all of the information contained in the section "Preparing Your Data Provider for a Parallel Connection with SAS" on page 69.

2 Make sure that you understand how the analytics environment configuration script works, as described in "How the Configuration Script Works" on page 82.

3 The software that is needed for the 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_5/Linux_for_x64`.

4 Copy the TKGrid_REP file that is appropriate for your operating system to the `/tmp` directory of the root node of the analytic cluster.

5 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 23.

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

7 Run the shell script in this directory.

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

8 Respond to the prompts from the configuration program:

Table 6.3 Configuration Parameters for the TKGrid_REP Shell Script

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 5.6 on page 72.)
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 5.3 on page 71.)
Do you want to configure remote access to Hadoop? (yes/NO)	If you are using a Hadoop machine cluster for your data provider, enter **y** and press Enter. Otherwise, enter **n** and press Enter.
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.

Parameter	Description
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 5.2 on page 70.)
Enter path of the directory containing the Hadoop and client jars.	Enter the path where the Cloudera Hadoop JAR files required by SAS reside. (This path was recorded earlier in Table 5.1 on page 70.)
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 5.4 on page 71.)
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 5.5 on page 72.)
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 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 6 on page 78.

Parameter	Description
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.

9 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.

10 You have finished deploying the analytics environment for a remote data source. If you have not done so already, install the appropriate SAS Embedded Process on the remote data appliance or 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/onlinedoc/indbtech/index.html**.

Validating the Analytics Environment Deployment

Overview of Validating

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

- “Use simsh to Validate” on page 88.
- “Use MPI to Validate” on page 88.

Use simsh to Validate

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:

```
/HPA-environment-installation-directory/bin/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 Provider” on page 62.

Use MPI to Validate

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-installation-directory/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 Provider” on page 62.

Appendix 1

Hadoop JAR Files Required for Remote Data Store Access

Overview of Hadoop JAR Files Required for Remote Data Store Access

Before you can configure the SAS High-Performance Analytics environment to use with a Hadoop not supplied by SAS, there are certain Hadoop JAR files that you must located on blade 0 of your SAS analytics cluster:

- “Cloudera Hadoop JAR Files” on page 92
- “Hortonworks Data Platform Hadoop” on page 94

Cloudera Hadoop JAR Files

Cloudera 4.5 Hadoop JAR files

The topic, "Prepare for Hadoop" on page 69 instructs you to copy certain Cloudera 4.5 JAR files to blade 0 on your analytics cluster. Co-locating these Cloudera JAR files on your analytics cluster, enables the SAS High-Performance Analytics environment to analyze data transferred from your Cloudera Hadoop cluster.

This section lists the Cloudera 4.5 JAR files that you need to copy. These JAR files reside on your Cloudera cluster in a directory under ***Cloudera-install-root*/parcels/CDH/lib/*Hadoop-service***.

Cloudera 4.5 Core Hadoop JAR files

```
activation-1.1.jar
asm-3.2.jar
avro-1.7.4.jar
cloudera-jets3t-2.0.0-cdh4.5.0.jar
commons-beanutils-1.7.0.jar
commons-beanutils-core-1.8.0.jar
commons-cli-1.2.jar
commons-codec-1.4.jar
commons-collections-3.2.1.jar
commons-compress-1.4.1.jar
commons-configuration-1.6.jar
commons-digester-1.8.jar
commons-el-1.0.jar
commons-httpclient-3.1.jar
commons-io-2.1.jar
commons-lang-2.5.jar
commons-logging-1.1.1.jar
commons-math-2.1.jar
commons-net-3.1.jar
guava-11.0.2.jar
```

```
hadoop-annotations-2.0.0-cdh4.5.0.jar
hadoop-auth-2.0.0-cdh4.5.0.jar
hadoop-common-2.0.0-cdh4.5.0.jar
hadoop-hdfs-2.0.0-cdh4.5.0.jar
hadoop-yarn-api-2.0.0-cdh4.5.0.jar
hadoop-yarn-client-2.0.0-cdh4.5.0.jar
hadoop-yarn-common-2.0.0-cdh4.5.0.jar
hadoop-yarn-server-common-2.0.0-cdh4.5.0.jar
hive-beeline-0.10.0-cdh4.5.0.jar
hive-builtins-0.10.0-cdh4.5.0.jar
hive-cli-0.10.0-cdh4.5.0.jar
hive-common-0.10.0-cdh4.5.0.jar
hive-contrib-0.10.0-cdh4.5.0.jar
hive-exec-0.10.0-cdh4.5.0.jar
hive-hbase-handler-0.10.0-cdh4.5.0.jar
hive-hwi-0.10.0-cdh4.5.0.jar
hive-jdbc-0.10.0-cdh4.5.0.jar
hive-metastore-0.10.0-cdh4.5.0.jar
hive-pdk-0.10.0-cdh4.5.0.jar
hive-serde-0.10.0-cdh4.5.0.jar
hive-service-0.10.0-cdh4.5.0.jar
hive-shims-0.10.0-cdh4.5.0.jar
hue-plugins-2.5.0-cdh4.5.0.jar
jackson-core-asl-1.8.8.jar
jackson-jaxrs-1.8.8.jar
jackson-mapper-asl-1.8.8.jar
jackson-xc-1.8.8.jar
jasper-compiler-5.5.23.jar
jasper-runtime-5.5.23.jar
jaxb-api-2.2.2.jar
jaxb-impl-2.2.3-1.jar
jersey-core-1.8.jar
jersey-json-1.8.jar
jersey-server-1.8.jar
jets3t-0.6.1.jar
jettison-1.1.jar
jetty-6.1.26.cloudera.2.jar
jetty-util-6.1.26.cloudera.2.jar
jline-0.9.94.jar
jsch-0.1.42.jar
jsp-api-2.1.jar
jsr305-1.3.9.jar
junit-4.8.2.jar
kfs-0.3.jar
log4j-1.2.17.jar
mockito-all-1.8.5.jar
paranamer-2.3.jar
```

```
protobuf-java-2.4.0a.jar
servlet-api-2.5.jar
slf4j-api-1.6.1.jar
snappy-java-1.0.4.1.jar
stax-api-1.0.1.jar
xmlenc-0.52.jar
xz-1.0.jar
```

Cloudera 4.5 Map Reduce 1 JAR files

```
hadoop-core-2.0.0-mr1-cdh4.5.0.jar
hadoop-tools-2.0.0-mr1-cdh4.5.0.jar
```

Note: Map Reduce 1 and Map Reduce 2 cannot be on the same Java class path.

Cloudera 4.5 Map Reduce 2 JAR files

```
hadoop-mapreduce-client-app-2.0.0-cdh4.5.0.jar
hadoop-mapreduce-client-common-2.0.0-cdh4.5.0.jar
hadoop-mapreduce-client-core-2.0.0-cdh4.5.0.jar
hadoop-mapreduce-client-jobclient-2.0.0-cdh4.5.0.jar
hadoop-mapreduce-client-shuffle-2.0.0-cdh4.5.0.jar
```

Note: Map Reduce 1 and Map Reduce 2 cannot be on the same Java class path.

Hortonworks Data Platform Hadoop

Overview of Hortonworks Data Platform Hadoop JAR Files

Choose the version of Hortonworks Data Platform (HDP) Hadoop that you plan to use with the SAS High-Performance Analytics environment:

- “Hortonworks Data Platform 1.3.2 JAR files” on page 94
- “Hortonworks Data Platform 2.0 JAR Files ” on page 96

Hortonworks Data Platform 1.3.2 JAR files

The topic, “Prepare for Hadoop” on page 69 instructs you to copy certain Hortonworks Data Platform 1.3.2 (HDP) JAR files to blade 0 on your analytics cluster. Co-locating

these HDP JAR files on your analytics cluster, enables the SAS High-Performance Analytics environment to analyze data transferred from your HDP Hadoop cluster.

This section lists the HDP 1.3.2 JAR files that you need to copy. These JAR files reside on your HDP cluster in a directory under **/usr/lib/hadoop** or **/usr/lib/hadoop/lib**.

Hortonworks 1.3.2 JAR files

```
ambari-log4j-1.2.5.17.jar
asm-3.2.jar
aspectjrt-1.6.11.jar
aspectjtools-1.6.11.jar
commons-beanutils-1.7.0.jar
commons-beanutils-core-1.8.0.jar
commons-cli-1.2.jar
commons-codec-1.4.jar
commons-collections-3.2.1.jar
commons-configuration-1.6.jar
commons-daemon-1.0.1.jar
commons-digester-1.8.jar
commons-el-1.0.jar
commons-httpclient-3.0.1.jar
commons-io-2.1.jar
commons-lang-2.4.jar
commons-logging-1.1.1.jar
commons-logging-api-1.0.4.jar
commons-math-2.1.jar
commons-net-3.1.jar
core-3.1.1.jar
guava-11.0.2.jar
hadoop-capacity-scheduler-1.2.0.1.3.2.0-111.jar
hadoop-client-1.2.0.1.3.2.0-111.jar
hadoop-core-1.2.0.1.3.2.0-111.jar
hadoop-fairscheduler-1.2.0.1.3.2.0-111.jar
hadoop-lzo-0.5.0.jar
hadoop-minicluster-1.2.0.1.3.2.0-111.jar
hadoop-thriftfs-1.2.0.1.3.2.0-111.jar
hadoop-tools-1.2.0.1.3.2.0-111.jar
hive-beeline-0.11.0.1.3.2.0-111.jar
hive-cli-0.11.0.1.3.2.0-111.jar
hive-common-0.11.0.1.3.2.0-111.jar
hive-contrib-0.11.0.1.3.2.0-111.jar
hive-exec-0.11.0.1.3.2.0-111.jar
hive-hbase-handler-0.11.0.1.3.2.0-111.jar
hive-hwi-0.11.0.1.3.2.0-111.jar
hive-jdbc-0.11.0.1.3.2.0-111.jar
```

```
hive-metastore-0.11.0.1.3.2.0-111.jar
hive-serde-0.11.0.1.3.2.0-111.jar
hive-service-0.11.0.1.3.2.0-111.jar
hive-shims-0.11.0.1.3.2.0-111.jar
hsqldb-1.8.0.10.jar
jackson-core-asl-1.8.8.jar
jackson-mapper-asl-1.8.8.jar
jasper-compiler-5.5.12.jar
jasper-runtime-5.5.12.jar
jdeb-0.8.jar
jersey-core-1.8.jar
jersey-json-1.8.jar
jersey-server-1.8.jar
jets3t-0.6.1.jar
jetty-6.1.26.jar
jetty-util-6.1.26.jar
jsch-0.1.42.jar
junit-4.5.jar
kfs-0.2.2.jar
log4j-1.2.15.jar
mockito-all-1.8.5.jar
netty-3.6.2.Final.jar
oro-2.0.8.jar
postgresql-9.1-901-1.jdbc4.jar
servlet-api-2.5-20081211.jar
slf4j-api-1.4.3.jar
slf4j-log4j12-1.4.3.jar
xmlenc-0.52.jar
```

Hortonworks Data Platform 2.0 JAR Files

The topic, “Prepare for Hadoop” on page 69 instructs you to copy certain Hortonworks Data Platform 2.0 (HDP) JAR files to blade 0 on your analytics cluster. Co-locating these HDP JAR files on your analytics cluster, enables the SAS High-Performance Analytics environment to analyze data transferred from your HDP Hadoop cluster.

This section lists the HDP 2.0 JAR files that you need to copy. These JAR files reside on your HDP cluster in a directory under **`/usr/lib/hadoop`** or **`/usr/lib/hadoop/lib`**.

Hortonworks Data Platform 2.0 JAR Files

```
activation-1.1.jar
ambari-log4j-1.4.3.38.jar
asm-3.2.jar
avro-1.7.4.jar
commons-beanutils-1.7.0.jar
commons-beanutils-core-1.8.0.jar
commons-cli-1.2.jar
commons-codec-1.4.jar
commons-collections-3.2.1.jar
commons-compress-1.4.1.jar
commons-configuration-1.6.jar
commons-digester-1.8.jar
commons-el-1.0.jar
commons-httpclient-3.1.jar
commons-io-2.1.jar
commons-lang-2.5.jar
commons-logging-1.1.1.jar
commons-math-2.1.jar
commons-net-3.1.jar
guava-11.0.2.jar
hadoop-annotations-2.2.0.2.0.6.0-101.jar
hadoop-auth-2.2.0.2.0.6.0-101.jar
hadoop-common-2.2.0.2.0.6.0-101.jar
hadoop-hdfs-2.2.0.2.0.6.0-101.jar
hadoop-hdfs-nfs-2.2.0.2.0.6.0-101.jar
hadoop-lzo-0.5.0.jar
hadoop-nfs-2.2.0.2.0.6.0-101.jar
hadoop-yarn-api-2.2.0.2.0.6.0-101.jar
hadoop-yarn-client-2.2.0.2.0.6.0-101.jar
hadoop-yarn-common-2.2.0.2.0.6.0-101.jar
hadoop-yarn-server-common-2.2.0.2.0.6.0-101.jar
hive-beeline-0.12.0.2.0.6.1-101.jar
hive-cli-0.12.0.2.0.6.1-101.jar
hive-common-0.12.0.2.0.6.1-101.jar
hive-contrib-0.12.0.2.0.6.1-101.jar
hive-exec-0.12.0.2.0.6.1-101.jar
hive-hbase-handler-0.12.0.2.0.6.1-101.jar
hive-hwi-0.12.0.2.0.6.1-101.jar
hive-jdbc-0.12.0.2.0.6.1-101.jar
hive-metastore-0.12.0.2.0.6.1-101.jar
hive-serde-0.12.0.2.0.6.1-101.jar
hive-service-0.12.0.2.0.6.1-101.jar
hive-shims-0.12.0.2.0.6.1-101.jar
jackson-core-asl-1.8.8.jar
jackson-jaxrs-1.8.8.jar
jackson-mapper-asl-1.8.8.jar
```

```
jackson-xc-1.8.8.jar
jasper-compiler-5.5.23.jar
jasper-runtime-5.5.23.jar
jaxb-api-2.2.2.jar
jaxb-impl-2.2.3-1.jar
jersey-core-1.9.jar
jersey-json-1.9.jar
jersey-server-1.9.jar
jets3t-0.6.1.jar
jettison-1.1.jar
jetty-6.1.26.jar
jetty-util-6.1.26.jar
jsch-0.1.42.jar
jsp-api-2.1.jar
jsr305-1.3.9.jar
junit-4.8.2.jar
log4j-1.2.17.jar
mockito-all-1.8.5.jar
mysql-connector-java.jar
netty-3.6.2.Final.jar
paranamer-2.3.jar
postgresql-9.1-901-1.jdbc4.jar
protobuf-java-2.5.0.jar
servlet-api-2.5.jar
slf4j-api-1.7.5.jar
slf4j-log4j12-1.7.5.jar
snappy-java-1.0.4.1.jar
stax-api-1.0.1.jar
xmlenc-0.52.jar
xz-1.0.jar
```

Hortonworks Data Platform 2.0 Map Reduce 2 JAR Files

```
hadoop-mapreduce-client-app-2.2.0.2.0.6.0-101.jar
hadoop-mapreduce-client-common-2.2.0.2.0.6.0-101.jar
hadoop-mapreduce-client-core-2.2.0.2.0.6.0-101.jar
hadoop-mapreduce-client-jobclient-2.2.0.2.0.6.0-101.jar
hadoop-mapreduce-client-shuffle-2.2.0.2.0.6.0-101.jar
```

Appendix 2

Updating the SAS High-Performance Analytics Infrastructure

Overview of Updating the Analytics Infrastructure

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

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

Update the 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 as the **root** user:

```
service sashpcmc stop
```

2 Update the management console using the following RPM command:

```
rpm -U /SAS-Software-Depot-Root-Directory/standalone_installs/
SAS_High-Performance_Computing_Management_Console/2_5/Linux_for_x64/
sashpcmc-2.5.x86_64.rpm
```

3 Log on to the console to validate your update.

Update Hadoop

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

1 Stop Hadoop by running the **/hadoop/hadoop/sbin/stop-dfs.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** You can issue a single **simsh** command to simultaneously check all the machines in the cluster: */HPA-environment-installation-directory*/bin/simsh ps –ef | grep hadoop.

3 Re-install Hadoop using **hadoopInstall** as described in "Deploying SAS High-Performance Deployment of Hadoop" on page 45.

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

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.

Update the Analytics Environment

You have the following options for managing updates to the SAS High-Performance Analytics environment:

- Delete the SAS High-Performance Analytics environment and install the newer version.

  See the procedure later in this topic.

- Rename the root installation directory for the current SAS High-Performance Analytics environment, and install the newer version under the previous root installation directory.

  See "Install the Analytics Environment" on page 78.

- Do nothing to the current SAS High-Performance Analytics environment, and install the new version under a new installation directory.

  See "Install the Analytics Environment" on page 78.

  When you change the path of the SAS High-Performance Analytics environment, you have to also have to reconfigure the SAS LASR Analytic Server to point to the new path. See "Add a SAS LASR Analytic Server" in Chapter 4 of *SAS Visual Analytics: Administration Guide*.

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 analytics environment processes running on any machine:

```
ps –ef | grep TKGrid
```

If you find any TKGrid processes, terminate them.

> **TIP** You can issue a single **simsh** command to simultaneously check all the machines in the cluster: */HPA-environment-installation-directory*/bin/simsh ps –ef | grep TKGrid.

2 Delete the analytics environment installation directory on every machine in the cluster:

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

> **TIP** You can issue a single **simsh** command to simultaneously remove the environment install directories on all the machines in the cluster: */HPA-environment-installation-directory*/bin/simsh rm -r -f */HPA-environment-installation-directory*.

3 Re-install the analytics environment using the shell script as described in “Install the Analytics Environment” on page 78.

Appendix 3

SAS High-Performance Analytics Infrastructure Command Reference

The **simsh** and **simcp** commands are installed with SAS High-Performance Computing Management Console and the SAS High-Performance Analytics environment. The default path to the commands is ***/HPCMC-installation-directory*/webmin/utilbin** and ***/HPA-environment-installation-directory*/bin**, respectively. Any user account that can access the commands and has passwordless secure shell configured can use them.

TIP Add one of the earlier referenced installation paths to your system PATH variable to make invoking **simsh** and **simcp** easier.

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:

```
/HPA-environment-installation-directory/bin/simcp /etc/hosts /etc
```

Appendix 4

SAS High-Performance Analytics Environment Client-Side Environment Variables

The following environment variables can be used on the client side to control the connection to the SAS High-Performance Analytics environment. You can set these environment variables in the following ways:

- invoke them in your SAS program using `options set=`
- add them to your shell before running the SAS program
- add them to your sasenv_local configuration file, if you want them used in all SAS programs

GRIDHOST=
: identifies the root node on the SAS High-Performance Analytics environment to which the client connects.

  The values for GRIDHOST and GRIDINSTALLLOC can both be specified in the GRIDHOST variable, separated by a colon (similar to the format used by **scp**). For example:

```
GRIDHOST=my_machine_cluster_001:/opt/TKGrid
```

GRIDINSTALLLOC=
: identifies the location on the machine cluster where the SAS High-Performance Analytics environment is installed. For example:

```
GRIDINSTALLLOC=/opt/TKGrid
```

GRIDMODE=SYM | ASYM
toggles the SAS High-Performance Analytics environment between symmetric (default) and asymmetric mode.

GRIDRSHCOMMAND= " " | " *ssh-path*"
(optional) specifies **rsh** or **ssh** used to launch the SAS High-Performance Analytics environment.

If unspecified or a null value is supplied, a SAS implementation of the SSH protocol is used.

ssh-path specifies the path to the SSH executable that you want to use. This can be useful in deployments where export controls restrict SAS from delivering software that uses cryptography. For example:

```
option set=GRIDRSHCOMMAND="/usr/bin/ssh";
```

GRIDPORTRANGE=
identifies the port range for the client to open. The root node connects back to the client using ports in the specified range. For example:

```
option set=GRIDPORTRANGE=7000-8000;
```

GRIDREPLYHOST=
specifies the name of the client machine to which the SAS High-Performance Analytics environment connects. GRIDREPLYHOST is used when the client has more than one network card or when you need to specify a full network name.

GRIDREPLYHOST can be useful when you need to specify a fully qualified domain name, when the client has more than one network interface card, or when you need to specify an IP address for a client with a dynamically assigned IP address that domain name resolution has not registered yet. For example:

```
GRIDREPLYHOST=myclient.example.com
```

Appendix 5

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 the Management Console

SELinux Modifications for the 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 the 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 Hadoop

SELinux Modifications for 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 Hadoop

The SAS High-Performance Deployment of 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 the Analytics Environment

SELinux Modifications for the 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 the 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 SAS LASR Analytic 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.

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
```

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

U