



SAS[®] High-Performance Analytics Infrastructure 3.5: Installation and Configuration Guide

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2016. *SAS® High-Performance Analytics Infrastructure 3.5: Installation and Configuration Guide*. Cary, NC: SAS Institute Inc.

SAS® High-Performance Analytics Infrastructure 3.5: Installation and Configuration Guide

Copyright © 2016, SAS Institute Inc., Cary, NC, USA

All Rights Reserved. Produced in the United States of America.

For a hard copy book: No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

For a web download or e-book: Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

U.S. Government License Rights; Restricted Rights: The Software and its documentation is commercial computer software developed at private expense and is provided with RESTRICTED RIGHTS to the United States Government. Use, duplication, or disclosure of the Software by the United States Government is subject to the license terms of this Agreement pursuant to, as applicable, FAR 12.212, DFAR 227.7202-1(a), DFAR 227.7202-3(a), and DFAR 227.7202-4, and, to the extent required under U.S. federal law, the minimum restricted rights as set out in FAR 52.227-19 (DEC 2007). If FAR 52.227-19 is applicable, this provision serves as notice under clause (c) thereof and no other notice is required to be affixed to the Software or documentation. The Government's rights in Software and documentation shall be only those set forth in this Agreement.

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

November 2016

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

3.5-P1:hpaicg

Contents

<i>What's New in Installation and Configuration for SAS High-Performance Analytics Infrastructure 3.5</i>	<i>v</i>
<i>Accessibility</i>	<i>vii</i>
Chapter 1 / Introduction to Deploying the SAS High-Performance Analytics Infrastructure	1
What Is Covered in This Document?	1
Which Version Do I Use?	2
What is the Infrastructure?	2
Where Do I Locate My Analytics Cluster?	4
Deploying the Infrastructure	9
Chapter 2 / Preparing Your System to Deploy the SAS High-Performance Analytics Infrastructure	13
Infrastructure Deployment Process Overview	13
System Settings for the Infrastructure	14
List the Machines in the Cluster or Appliance	15
Review Passwordless Secure Shell (SSH) Requirements	16
Preparing for Kerberos	17
Preparing to Install SAS High-Performance Computing Management Console	20
Requirements for Co-located Hadoop	21
Preparing to Deploy the SAS High-Performance Analytics Environment	21
Recommended Database Names	23
Pre-installation Ports Checklist for SAS	23
Chapter 3 / Deploying SAS High-Performance Computing Management Console	25
Infrastructure Deployment Process Overview	25
Benefits of the Management Console	26
Overview of Deploying the Management Console	26
Installing the Management Console	27
Configure the Management Console	28
Create the Installer Account and Propagate the SSH Key	29
Create the First User Account and Propagate the SSH Key	32
Chapter 4 / Modifying Co-located Hadoop	37
Infrastructure Deployment Process Overview	37
Overview of Modifying Co-located Hadoop	38
Modify the Existing Apache Hadoop Cluster, Version 2.7	39
Modify the Existing Cloudera Hadoop Cluster, Version 5	41
Modify the Existing Cloudera Hadoop Cluster, Version 4	44
Modify the Existing Hortonworks Data Platform Hadoop Cluster	46
Modify the Existing IBM BigInsights Hadoop Cluster	48
Modify the Existing MapR Hadoop Cluster	50
Modify the Existing Pivotal HD Hadoop Cluster	50
Hadoop Configuration Step for SAS Visual Analytics	52
Chapter 5 / Deploying the SAS High-Performance Analytics Environment	53
Infrastructure Deployment Process Overview	53

Overview of Deploying the Analytics Environment	54
Encrypting SASHDAT Files	57
Install the Analytics Environment	58
Configuring the Analytics Environment for SASHDAT Encryption	61
Validating the Analytics Environment Deployment	62
Resource Management for the Analytics Environment	64
Chapter 6 / Configuring the Analytics Environment for a Remote Parallel Connection	67
Infrastructure Deployment Process Overview	67
Overview of Configuring the Analytics Environment for a Remote Parallel Connection	68
Preparing for a Remote Parallel Connection	69
Configuring for Access to a Data Store with a SAS Embedded Process	75
Appendix 1 / Updating the SAS High-Performance Analytics Infrastructure	79
Analytics Infrastructure Dependencies	79
Updating Co-located Hadoop	79
Update the Analytics Environment	83
Updating the SAS High-Performance Computing Management Console	84
Appendix 2 / SAS High-Performance Analytics Infrastructure Command Reference	85
Appendix 3 / SAS High-Performance Analytics Environment Client-Side Environment Variables	87
Appendix 4 / Deploying on SELinux and IPTables	89
Overview of Deploying on SELinux and IPTables	89
Prepare the Management Console	90
Prepare the Analytics Environment	90
Analytics Environment Post-Installation Modifications	91
iptables File	91
Appendix 5 / Setting Up Passwordless Secure Shell (SSH)	93
Recommended Reading	97
Glossary	99
Index	107

What's New

What's New in Installation and Configuration for SAS High-Performance Analytics Infrastructure 3.5

Overview

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

- SAS High-Performance Computing Management Console 2.8
- SAS Plug-ins for Hadoop, version 1.0
- SAS High-Performance Analytics environment 3.5
(also referred to as the SAS High-Performance Node Installation)

SAS High-Performance Analytics Infrastructure 3.5 includes the following changes and enhancements:

- [“SAS High-Performance Deployment for Hadoop Discontinued”](#) on page v
- [“SAS Plug-ins for Hadoop”](#) on page vi

SAS High-Performance Deployment for Hadoop Discontinued

SAS has discontinued SAS High-Performance Deployment for Hadoop.

If you want to co-locate the SAS High-Performance Analytics environment with Hadoop, you can select and install a supported commercial version of Hadoop that meets your business requirements. Or you can choose to download and deploy an Apache Hadoop distribution.

Supported Hadoop distributions that you co-locate with the analytics environment must be modified with SAS Plug-ins for Hadoop.

For more information, see [“Overview of Modifying Co-located Hadoop” on page 38](#).

SAS Plug-ins for Hadoop

SAS Plug-ins for Hadoop is a package that provides services to your supported Hadoop distribution. These services enable the SAS High-Performance Analytics environment to write SASHDAT file blocks evenly across the HDFS file system. For more information, see [“Overview of Modifying Co-located Hadoop” on page 38](#).

Accessibility

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

1

Introduction to Deploying the SAS High-Performance Analytics Infrastructure

<i>What Is Covered in This Document?</i>	1
<i>Which Version Do I Use?</i>	2
<i>What is the Infrastructure?</i>	2
<i>Where Do I Locate My Analytics Cluster?</i>	4
Overview of Locating Your Analytics Cluster	4
Analytics Cluster Co-Located with Your Hadoop Cluster	5
Analytics Cluster Remote from Your Data Store (Serial Connection)	6
Analytics Cluster Remote from Your Data Store (Parallel Connection)	7
Hadoop Deployment Comparison	8
<i>Deploying the Infrastructure</i>	9
Overview of Deploying the Infrastructure	9
Step 1: Create a SAS Software Depot	9
Step 2: Check for Documentation Updates	9
Step 3: Prepare Your Analytics Cluster	10
Step 4: (Optional) Deploy SAS High-Performance Computing Management Console	10
Step 5: (Optional) Modify Co-located Hadoop	10
Step 6: Deploy the SAS High-Performance Analytics Environment	10
Step 7: (Optional) Deploy the SAS Embedded Process for Hadoop	10
Step 8: (Optional) Configure the Analytics Environment for a Remote Parallel Connection	11

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 Plug-ins for 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 components:

- SAS High-Performance Computing Management Console, version 2.8
- SAS Plug-ins for Hadoop, version 1.0
- SAS High-Performance Analytics environment, version 3.5

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

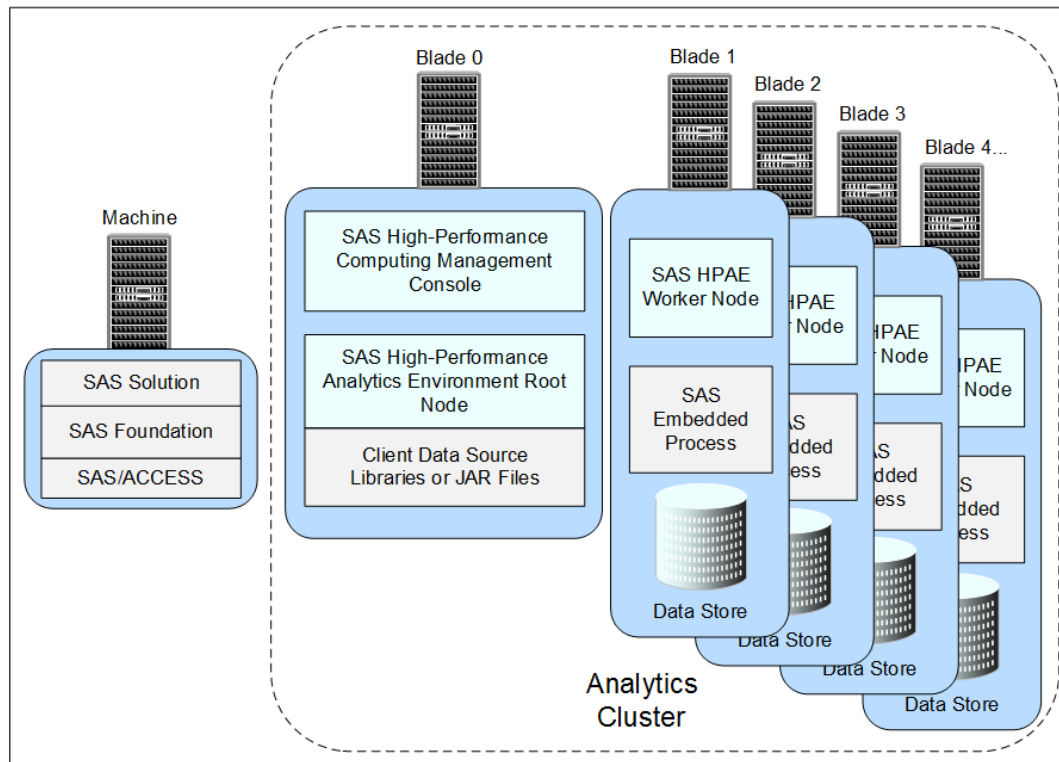
Refer to your order summary to determine the specific version of the infrastructure that is included in your SAS order. Your order summary resides in your SAS Software Depot for your respective order under the `install_doc` directory (for example, `C:\SAS Software Depot\install_doc\my-order\ordersummary.html`).

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 analytics cluster. The analytics cluster is a Hadoop cluster or a data appliance.

- (Optional) SAS Plug-ins for Hadoop

Some solutions, such as SAS Visual Analytics, rely on a SAS data store that is co-located with the SAS High-Performance Analytics environment on the analytics cluster. One option for this co-located data store is SAS Plug-ins for Hadoop.

If you already have one of the supported Hadoop distributions, you can modify it with files from the SAS Plug-ins for Hadoop package. Hadoop modified with SAS Plug-ins for Hadoop enables the SAS High-Performance Analytics environment to write SASHDAT file blocks evenly across the HDFS file system. This even distribution provides a balanced workload across the machines in the cluster and enables SAS analytic processes to read SASHDAT tables at very impressive rates.

For more information, see [“Overview of Modifying Co-located Hadoop” on page 38](#).

- (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 analytics cluster are all made easier by the management console.

Other software on the analytics cluster includes 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 analytics 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](#) and the [SAS/ACCESS for Relational Databases: Reference](#).

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](#)
- ☐ [SAS Customer Intelligence](#)
- ☐ [SAS High-Performance Risk](#)
- ☐ [SAS Visual Analytics](#)

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 analytics cluster placement:

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

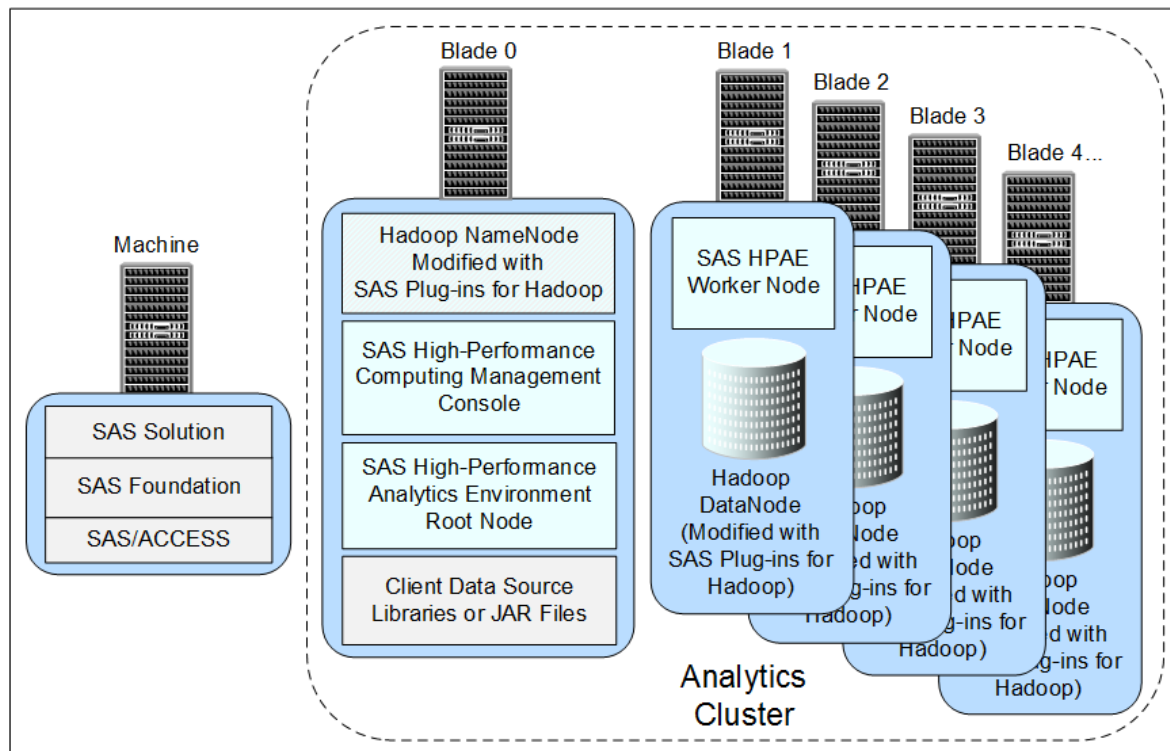
TIP Where you locate your cluster depends on a number of criteria. Your SAS representative will know the latest supported configurations, and can work with you to help you determine which cluster placement option works best for your site. Also, there might be solution-specific criteria that you should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

Analytics Cluster Co-Located with Your Hadoop Cluster

Note: In a co-located configuration, the SAS High-Performance Analytics environment supports the Apache, Cloudera, Hortonworks, IBM BigInsights, MapR, and Pivotal HD distributions of Hadoop. For more specific version information, see the [SAS 9.4 Supported Hadoop Distributions](#).

The following figure shows the analytics cluster co-located on your Hadoop cluster:

Figure 1.2 Analytics Cluster Co-Located with the Hadoop Cluster

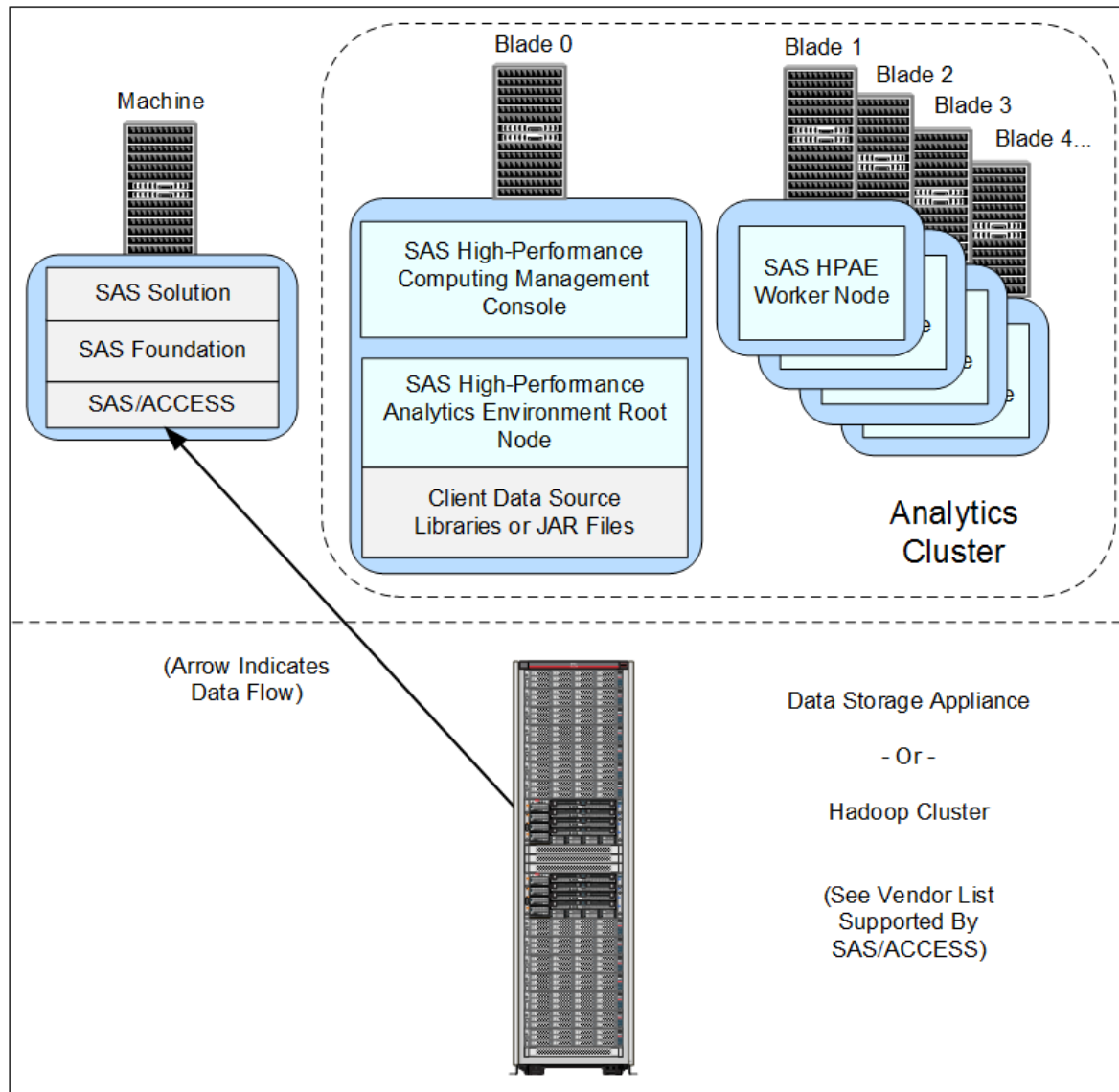


Note: For deployments that use Hadoop for the co-located data provider and access SASHDAT tables exclusively, SAS/ACCESS and the SAS Embedded Process are not needed.

Analytics 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)



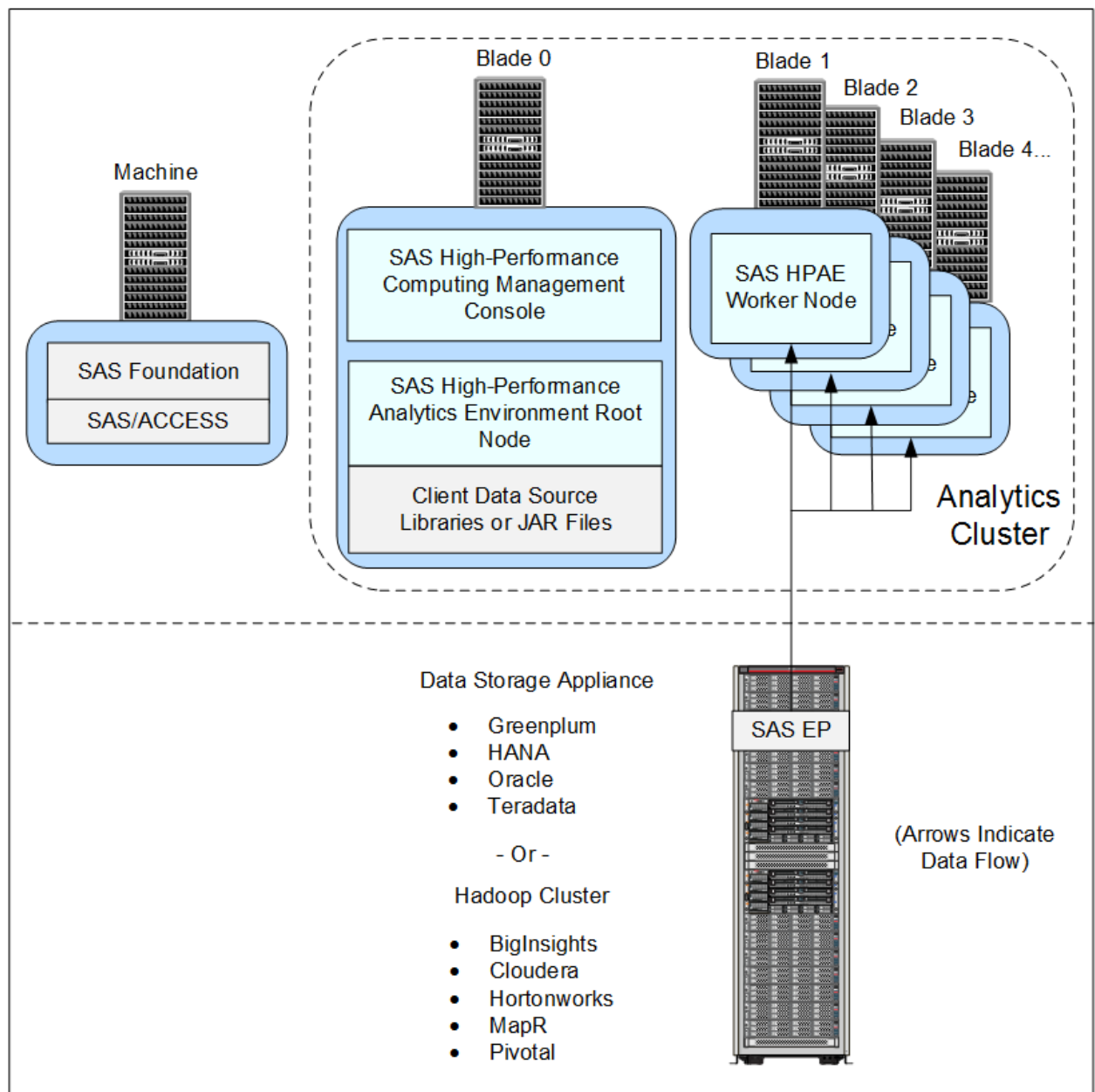
The serial connection between the analytics cluster and your data store is achieved by using the SAS/ACCESS Interface. SAS/ACCESS is orderable in a deployment package that is specific for your data source. For more information, refer to the [SAS/ACCESS for Relational Databases: Reference](#).

Analytics Cluster Remote from Your Data Store (Parallel Connection)

Note: In the third maintenance release of SAS 9.4, SAS Embedded Process supports the Cloudera, Hortonworks, IBM BigInsights, MapR, and Pivotal HD distributions of Hadoop. For more specific version information, see the [SAS 9.4 Support for Hadoop](#).

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 analytics 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](#).

Hadoop Deployment Comparison

The following table compares various deployment Hadoop scenarios.

Table 1.1 Hadoop Deployment Comparison

	Co-located with Hadoop SASHDAT: Yes SAS Embedded Process: No	Co-located with Hadoop SASHDAT: No SAS Embedded Process: Yes	Remote Data Provider SASHDAT: Not Supported SAS Embedded Process: No	Remote Data Provider SASHDAT: Not Supported SAS Embedded Process: Yes
SASHDAT Support	Yes	No	No, SASHDAT is co-located or MapR NFS only.	No. SASHDAT is co-located or MapR NFS only.
Parallel R/W	Yes for SASHDAT and CSV. No for SAS/ACCESS because there is no SAS Embedded Process.	Yes. (At least for PROC HDMD.)	No, SAS/ACCESS can perform a serial read through the root node.	Yes. SAS/ACCESS and SAS Embedded Process enable this.
Asymmetric*	No for SASHDAT. No for SAS/ACCESS. SAS/ACCESS can perform a serial read through the root node.	Yes. SAS Embedded Process on all the machines can deliver data to a fewer or greater number of machines.	No. SAS/ACCESS can perform a serial read through the root node.	Yes. SAS Embedded Process on all the machines can deliver data to a fewer or greater number of machines.
Serial Reads for SAS/ACCESS	SAS/ACCESS reads are always serial without SAS Embedded Process.	(If something is misconfigured, SAS/ACCESS performs a serial read.)	SAS/ACCESS reads are always serial without SAS Embedded Process.	(If something is misconfigured, SAS/ACCESS performs a serial read.)
Popularity	This is the SAS Visual Analytics configuration.	Rare.	Rare.	Popular. (Can be combined with a co-located Hadoop configuration.)

* *Asymmetric* refers to a deployment where the total number of SAS High-Performance Analytics environment worker nodes is *not* equal to the total number of Hadoop data nodes. *Symmetric* refers to an equal number of worker nodes and data nodes.

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) Modify co-located Hadoop.
6. Deploy the SAS High-Performance Analytics environment.
7. (Optional) Deploy the SAS Embedded Process for Hadoop.
8. (Optional) Configure the analytics environment for a remote parallel connection

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 chosen 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 *SAS Intelligence Platform: Installation and Configuration Guide*.

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](#).

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 13.

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.

For more information, see [Chapter 3, “Deploying SAS High-Performance Computing Management Console,”](#) on page 25.

Step 5: (Optional) Modify Co-located Hadoop

If your site wants to use Hadoop as the co-located data store, then you can modify a supported pre-existing Hadoop distribution.

For more information, see [Chapter 4, “Modifying Co-Located Hadoop ,”](#) on page 37.

Step 6: Deploy the SAS High-Performance Analytics Environment

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

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 5, “Deploying the SAS High-Performance Analytics Environment,”](#) on page 53.

Step 7: (Optional) Deploy the SAS Embedded Process for Hadoop

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 analytics

cluster. These components are contained in a deployment package that is specific for your data source.

For information about installing the SAS Embedded Process, see the [SAS In-Database Products: Administrator's Guide](#).

Step 8: (Optional) Configure the Analytics Environment for a Remote Parallel Connection

You can optionally configure the SAS High-Performance Analytics Environment for a remote parallel connection.

For more information, see [Chapter 6, "Configuring the Analytics Environment for a Remote Parallel Connection,"](#) on page 67.

2

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

<i>Infrastructure Deployment Process Overview</i>	13
<i>System Settings for the Infrastructure</i>	14
<i>List the Machines in the Cluster or Appliance</i>	15
<i>Review Passwordless Secure Shell (SSH) Requirements</i>	16
<i>Preparing for Kerberos</i>	17
Kerberos Prerequisites	17
Generate and Test Host Principals: Example	17
Configure Passwordless SSH to Use Kerberos	18
Preparing the Analytics Environment for Kerberos	19
<i>Preparing to Install SAS High-Performance Computing Management Console</i>	20
User Account Considerations for the Management Console	20
Management Console Requirements	20
<i>Requirements for Co-located Hadoop</i>	21
<i>Preparing to Deploy the SAS High-Performance Analytics Environment</i>	21
User Accounts for the SAS High-Performance Analytics Environment	21
Consider Umask Settings	22
Additional Prerequisite for Greenplum Deployments	23
<i>Recommended Database Names</i>	23
<i>Pre-installation Ports Checklist for SAS</i>	23

Infrastructure Deployment Process Overview

Preparing your analytics cluster is the third of eight 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) Modify co-located Hadoop.
- 6. Deploy the SAS High-Performance Analytics environment.
- 7. (Optional) Deploy the SAS Embedded Process for Hadoop.
- 8. (Optional) Configure the analytics environment for a remote parallel connection.

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:

- Edit the `/etc/sudoers` file to disable the requirement for a terminal:


```
Defaults !requiretty
```

Also, when issuing a `sudo` command with the simultaneous commands (`simcp`, `simsh`) be sure to place `sudo` first. For example:

```
sudo /opt/TKGrid/bin/simsh ls -dl /root
```

These steps address a known bug in Red Hat Linux:
https://bugzilla.redhat.com/show_bug.cgi?id=1020147
 - 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 1.1 GB of disk space. 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](#).

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 on Greenplum), create an `/etc/gridhosts` file for use by SAS High-Performance Computing Management Console, SAS Plug-ins for Hadoop, and the SAS High-Performance Analytics environment. (The grid hosts file is copied to the other machines in the cluster during the installation process.) For more information, see [Chapter 3, “Deploying SAS High-Performance Computing Management Console,”](#) on page 25 before you start the installation.

CAUTION! The `gridhosts` file must contain only those machines that are members of your analytics cluster or data appliance. These machines are the NameNode (or Root Node) and its DataNodes (or Worker Nodes). If the management console is located on a machine that is not a member of the analytics cluster, then this machine must also contain a copy of `/etc/gridhosts` with its host name added to the list of machines.

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.

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

- [Supported Hadoop distributions](#): NameNode (blade 0)
- Greenplum Data Computing Appliance: Master Server

The following lines are an example of the file contents:

```
grid001
grid002
grid003
grid004
...
```

Note: Care must be taken with your `gridhosts` file. The `gridhosts` file can appear to be readable, but the SAS High-Performance Analytics environment can skip entries when whitespace characters (such as tabs) have been inadvertently inserted.

TIP You can use SAS High-Performance Computing Management Console to create and manage your grid hosts file. For more information, see the [SAS High-Performance Computing Management Console: User's Guide](#).

Review Passwordless Secure Shell (SSH) Requirements

TIP If you are not familiar with passwordless Secure Shell (SSH) please see [Appendix 5, “Setting Up Passwordless Secure Shell \(SSH\),” on page 93.](#)

Secure Shell (SSH) has the following requirements:

- To support Kerberos, enable GSSAPI authentication methods in your implementation of Secure Shell (SSH).

Note: If you are using Kerberos, see [“Configure Passwordless SSH to Use Kerberos” on page 18](#).

- Passwordless Secure Shell (SSH) is required [on all machines in the cluster or on the data appliance.](#)

Note the following items:

- If your SAS compute server is on a separate system from the SAS High-Performance Analytics environment name node, then you need to copy the `~/.ssh` directory on the compute server machine as well.
- Passwordless SSH must be bi-directional between all of the analytics environment worker nodes.
- The following user accounts require passwordless SSH:
 - 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 20.](#)
 - Hadoop user account

For more information about Hadoop user accounts, see [“Requirements for Co-located Hadoop” on page 21.](#)
 - SAS High-Performance Analytics environment user account

Passwordless SSH must be configured for every user that runs a SAS High-Performance procedure or interacts with SAS LASR Analytic Server or SAS data (sashdat format).

For more information about the environment’s user accounts, see [“Preparing to Deploy the SAS High-Performance Analytics Environment” on page 21.](#)

TIP Users’ home directories must be located in the same directory on each machine in the analytics cluster. For example, you will experience problems if user foo has a home directory at `/home/foo` on blade one and blade two, and a home directory at `/mnt/user/foo` on blade three.

Preparing for Kerberos

Kerberos Prerequisites

The SAS High-Performance Analytics infrastructure supports the Kerberos computer network authentication protocol. Throughout this document, we indicate the particular settings that you need to perform in order to make parts of the infrastructure configurable for Kerberos. However, you must understand and be able to verify your security setup. If you are using Kerberos, you need the ability to get a Kerberos ticket.

The list of Kerberos prerequisites is:

- A Kerberos key distribution center (KDC)
- All machines configured as Kerberos clients
- Permissions to copy and secure Kerberos keytab files on all machines
- A user principal for the Hadoop user
(This is used for setting up the cluster and performing administrative functions.)
- Encryption types supported on the Kerberos domain controller should be aes256-cts:normal and aes128-cts:normal

Generate and Test Host Principals: Example

This topic provides an example of setting up hosts using MIT Kerberos. There are other implementations of Kerberos, such as Microsoft Active Directory, that the SAS High-Performance Analytics infrastructure supports.

Every machine in the analytics cluster must have a host principal and a Kerberos keytab in order to operate as Kerberos clients.

To generate and test host principals, follow these steps:

- 1 Execute `kadmin.local` on the KDC.
- 2 Run the following command for each machine in the cluster:

```
addprinc -randkey +ok_to_delegate host/$machine-name
```

where *machine-name* is the host name of the particular machine.
- 3 Generate host keytab files in `kadmin.local` for each machine, by running the following command:

```
ktadd -norandkey -k $machine-name.keytab host/$machine-name
```

where *machine-name* is the name of the particular machine.

TIP When generating keytab files, it is a best practice to create files by machine. In the event a keytab file is compromised, the keytab will contain only the host principal associated with machine it resides on, instead of a single file that contains every machine in the environment.

- 4 Copy each generated keytab file to its respective machine under `/etc`, rename the file to `krb5.keytab`, and secure it with mode 600 and owned by root.

For example:

```
cp keytab /etc/krb5.keytab
chown root:root /etc/krb5.keytab
chmod 600 /etc/krb5.keytab
```

- 5 Validate your configuration in a temporary credential cache (ccache) to avoid overwriting any ccache in your user session with the host's credentials:

```
kinit -kt /etc/krb5.keytab -c ~/testccache host/
machine.name@REALM.NAME
```

- 6 Because kinit obtains only a krbtgt ticket for a given principal, also validate that Kerberos is able to issue service tickets for the host principal:

```
kvno -c ~/testccache machine.name@REALM.NAME
```

- 7 Run the klist command to check the status of your Kerberos ticket:

```
klist -efac ~/testccache
```

Your klist output should resemble the following:

```
Ticket cache: FILE:/home/myacct/testccache
Default principal: host/myserver.example.com@NA.EXAMPLE.COM

Valid starting    Expires          Service principal
07/07/15 15:33:32 07/08/15 01:33:32 krbtgt/NA.EXAMPLE.COM@NA.EXAMPLE.COM
    renew until 07/14/15 15:33:32, Flags: FRIA
    Etype (skey, tkt): aes256-cts-hmac-sha1-96, aes256-cts-hmac-sha1-96
    Addresses: (none)
07/07/15 15:34:09 07/08/15 01:33:32 host/myserver.example.com@NA.EXAMPLE.COM
    renew until 07/14/15 15:33:32, Flags: FRAO
    Etype (skey, tkt): aes256-cts-hmac-sha1-96, aes256-cts-hmac-sha1-96
    Addresses: (none)
```

Note: If you intend to deploy the SAS Embedded Process on the cluster for use with SAS/ACCESS Interface to Hadoop, then a user keytab file for the user ID that runs HDFS is required.

- 8 Delete your ccache:

```
kdestroy -c ~/testccache
```

Configure Passwordless SSH to Use Kerberos

TIP If you are not familiar with passwordless Secure Shell (SSH), please see [Appendix 5, “Setting Up Passwordless Secure Shell \(SSH\),” on page 93.](#)

Passwordless access of some form is a requirement of the SAS High-Performance Analytics environment through its use of the Message Passing Interface (MPI). Traditionally, public key authentication in Secure Shell (SSH) is used to meet the passwordless access requirement. For Secure Mode Hadoop, GSSAPI with Kerberos is used as the passwordless SSH mechanism. GSSAPI

with Kerberos not only meets the passwordless SSH requirements, but also supplies Hadoop with the credentials required for users to perform operations in HDFS with SAS LASR Analytic Server and SASHDAT files. Certain options must be set in the SSH daemon and SSH client configuration files. Those options are as follows and assume a default configuration of `sshd`.

To configure passwordless SSH to use Kerberos, follow these steps:

- 1 In the `sshd_config` file, set:

```
GSSAPIAuthentication yes
```

- 2 In the `ssh_config` file, set:

```
Host *.domain.net
```

```
GSSAPIAuthentication yes
```

```
GSSAPIDelegateCredentials yes
```

where *domain.net* is the domain name used by the machine in the cluster.

TIP Although you can specify `host *`, this is not recommended because it would allow GSSAPI Authentication with any host name.

Preparing the Analytics Environment for Kerberos

During start-up, the Message Passing Interface (MPI) sends a user's Kerberos credentials cache (KRB5CCNAME) that can cause an issue when Hadoop attempts to use Kerberos credentials to perform operations in HDFS.

Under Secure Shell (SSH), a random set of characters are appended to the credentials cache file, so the value of the KRB5CCNAME environment variable is different for each machine. To set the correct value for KRB5CCNAME on each machine, you must use the option below when asked for additional options to MPIRUN during the analytics environment installation:

```
-genvlist `env | sed -e s/=.*// | sed /KRB5CCNAME/d | tr -d '\n'`TKPATH,LD_LIBRARY_PATH
```

Note: Enter the above option on one line. Do not add any carriage returns or other whitespace characters.

For more information, see [Table 5.2 on page 59](#).

You must use a launcher that supports GSSAPI authentication because the implementation of SSH that is included with SAS does not support it. Add the following to your SAS programs on the client:

```
option set=GRIDRSHCOMMAND="/path-to-file/ssh";
```

TIP Adding GRIDRSHCOMMAND to your `sasv9_usermods.cfg` preserves the setting during SAS upgrades and avoids having to manually set that environment variable on the client before starting SAS.

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.

Note: The management console can manage operating system user accounts if the machines are configured to use the `/etc/passwd` local database only.

- 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 15](#).
- 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*.)

Requirements for Co-located Hadoop

If you already have one of the supported Hadoop distributions, you can modify it with files from the SAS Plug-ins for Hadoop package. Hadoop modified with SAS Plug-ins for Hadoop enables the SAS High-Performance Analytics environment to write SASHDAT file blocks evenly across the HDFS file system.

The following is required for existing Hadoop clusters with which the SAS High-Performance Analytics environment can be co-located:

- [Your Hadoop distribution must be supported.](#)
- Each machine in the cluster must be able to resolve the host name of all the other machines.
- The NameNode host does not also have a DataNode configured on it.
- These Hadoop directories must reside on local storage:
 - the directory on the file system where the Hadoop NameNode stores the namespace and transactions logs persistently
 - the directory on the file system where temporary MapReduce data is written
 - the directory on the file system where the MapReduce framework writes system files

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

- Time must be synchronized across all machines in the cluster.
- (Cloudera 5 only) Make sure that all machines configured for the SAS High-Performance Analytics environment are in the same role group.
- 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.

Preparing to Deploy the SAS High-Performance Analytics Environment

If you are using Kerberos, see also [“Preparing for Kerberos” on page 17](#).

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.

TIP 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 20](#).

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. Any user accounts running analytics environment processes must also be configured with passwordless SSH.

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.

Note: Remember that the LASRMonitor process that is part of SAS Visual Analytics must be run with an account (by default, sas) that can read the server signature file. (This signature file is created when you start a SAS LASR Analytic Server and the file is specified in SAS metadata. For more information, see [Establishing Connectivity to a SAS LASR Analytic Server](#) in the *SAS Intelligence Platform: Data Administration Guide*.)

You can also add umask settings to the resource settings file for the SAS Analytics environment. For more information, see [“Resource Management for the Analytics Environment” on page 64](#).

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 deploy the appropriate SAS/ACCESS interface and SAS Embedded Process that SAS supplies with SAS In-Database products. For more information, see the [SAS In-Database Products: Administrator's Guide](#).

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.

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 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 2.1 Pre-installation Checklist for SAS Ports

SAS Component	Default Port	Data Direction	Actual Port
SAS High-Performance Computing Management Console server	10020	Inbound	

SAS Component	Default Port	Data Direction	Actual Port
SAS Plug-in on the NameNode	15452	Inbound	
SAS Plug-in on the DataNode	15453	Inbound	

3

Deploying SAS High-Performance Computing Management Console

<i>Infrastructure Deployment Process Overview</i>	25
<i>Benefits of the Management Console</i>	26
<i>Overview of Deploying the Management Console</i>	26
<i>Installing the Management Console</i>	27
Install SAS High-Performance Computing Management Console Using RPM	27
Install the Management Console Using tar	27
<i>Configure the Management Console</i>	28
<i>Create the Installer Account and Propagate the SSH Key</i>	29
<i>Create the First User Account and Propagate the SSH Key</i>	32

Infrastructure Deployment Process Overview

Installing and configuring SAS High-Performance Computing Management Console is an optional fourth of eight 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) Modify co-located Hadoop.
6. Deploy the SAS High-Performance Analytics environment.
7. (Optional) Deploy the SAS Embedded Process for Hadoop.
8. (Optional) Configure the analytics environment for a remote parallel connection.

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 Analytics Server, passwordless SSH is required to run jobs on the machines in the cluster.

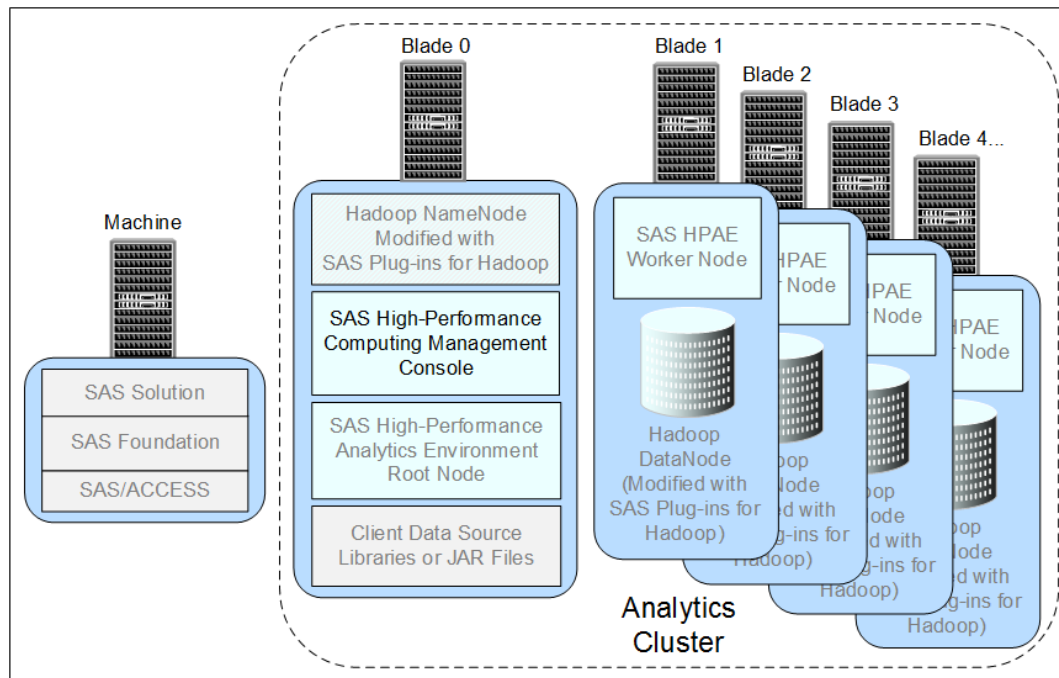
Also, users of some SAS solutions must have an operating system (external) account on all the machines in the cluster and must have the key distributed across the cluster. For more information, see [“Create the First User Account and Propagate the SSH Key”](#) on page 32.

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. For more information, see the [SAS High-Performance Computing Management Console: User's Guide](#).

Overview of Deploying the Management Console

The SAS High-Performance Computing Management Console is deployed on the machine where the SAS High-Performance Analytics environment is deployed. In this document, that machine is blade 0.

Figure 3.1 Management Console Deployed with a Hadoop Cluster



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 [Appendix 1, “Updating the SAS High-Performance Analytics Infrastructure,”](#) on page 79.

- 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 20.
- 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_8/Linux_for_x64` directory.
- 4 Enter one of the following commands:
 - To install in the default location of `/opt`:

```
rpm -ivh sashpcmc*.rpm
```
 - To install in a location of your choice:

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

where *directory* is an absolute path where you want to install the console.
- 5 Proceed to the topic [“Configure the Management Console”](#) on page 28.

Install the Management Console Using tar

Some versions of Linux use different RPM libraries and require an 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 20.
- 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_8/Linux_for_x64` directory.
- 4 Copy `sashpcmc-2.8.tar.gz` to the location where you want to install the management console.

- 5 Change to the directory where you copied the tar file, and run the following command:

```
tar -xzf sashpcmc-2.8.tar.gz
```

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

- 6 Proceed to the topic [“Configure the Management Console” on page 28](#).

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:

```
management-console-installation-directory/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 using system authentication

Use SSL\HTTPS (yes|no)

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

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

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

- 6 Make one of two choices:

- Enter **create** 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.

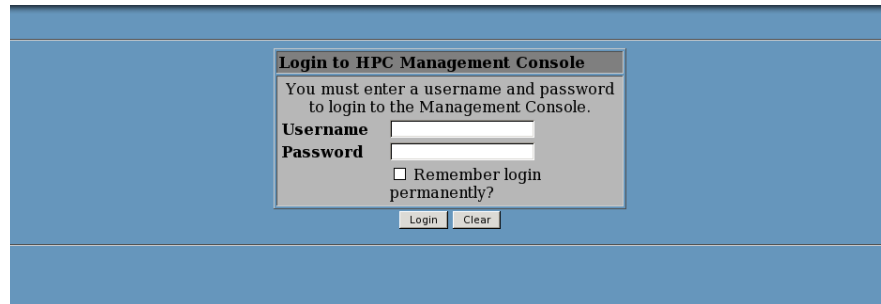
- 7 To start the SAS High-Performance Computing Management Console server, enter the following command from any directory:

```
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.

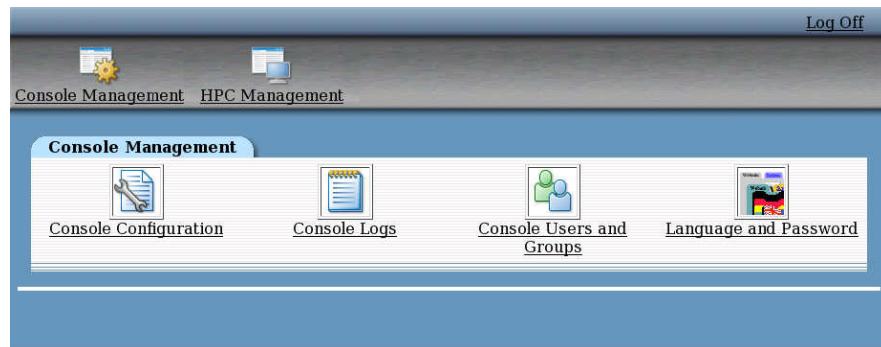
For example: `https://myserver.example.com:10020`

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:

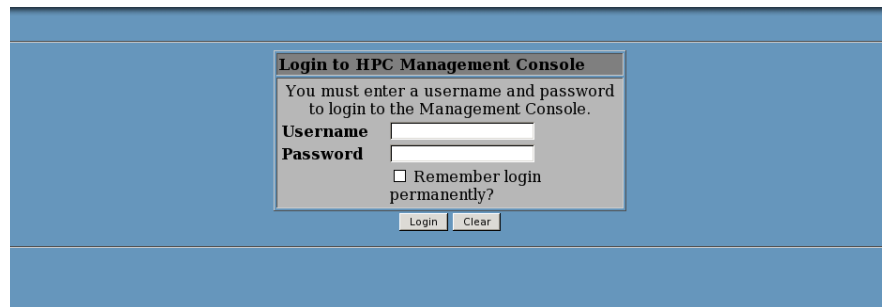
```
service sashpcmc status
```

(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 28](#).

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

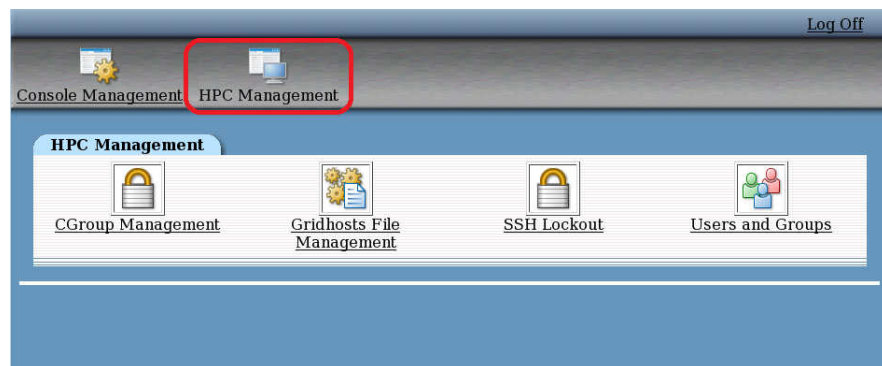
For example: `http://myserver.example.com:10020`

The Login page appears.



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

The Console Management page appears.



- 4 Click **HPC Management**.

The HPC Management page appears.



- 5 Click **Users and Groups**.

The Users and Groups page appears.

Log Off

Console Management HPC Management

Users and Groups

HPC Users HPC Groups Midtier Shared Key

Select all. | Invert selection. **Create a new user.**

	Username	User ID	Group	Real name	Home directory	Shell	SSH Keys?	Last login
<input type="checkbox"/>	sas	32237	sas		/home/sas	/bin/bash	Yes	
<input type="checkbox"/>	hadoop	32242	hadoop		/home/hadoop	/bin/bash	Yes	

Select all. | Invert selection. | Create a new user.

Delete Selected Users

Display Logins By ☒ All users ☐ Only user Show recent logins by all Unix users who have connected via SSH.

Show Logged in Users Show users who are currently logged in via SSH.

[Return to index](#)

6 Click **Create a new user**.

The Create User page appears.

Create User

User Details

Username

User ID ☒ Automatic ☐ Calculated

Real name

Home directory ☒ Automatic ☐ Directory

Shell

Password ☐ No password required ☐ Normal password ☐ Pre-encrypted password

Password Options

Password changed Expiry date

Minimum days Maximum days

Warning days Inactive days

Group Membership

Primary group ☒ New group with same name as user ☐ New group ☐ Existing group

Secondary groups **All groups** **In groups**

HPC Actions and Settings

Propagate User ☒ Yes ☐ No

Generate and Propagate SSH Keys ☒ Yes ☐ No

Add Shared Midtier Key ☐ Yes ☒ No

Upon Creation..

Create home directory? ☒ Yes ☐ No

Copy template files to home directory? ☒ Yes ☐ No

Create

[Return to users and groups list](#)

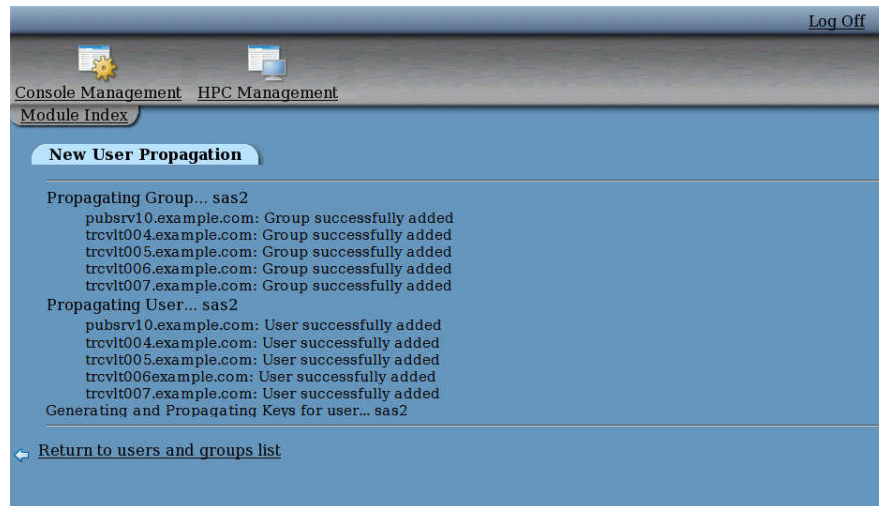
7 Enter information for the new user, using the security policies in place at your site.

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

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

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

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



Create the First User Account and Propagate the SSH Key

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

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

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

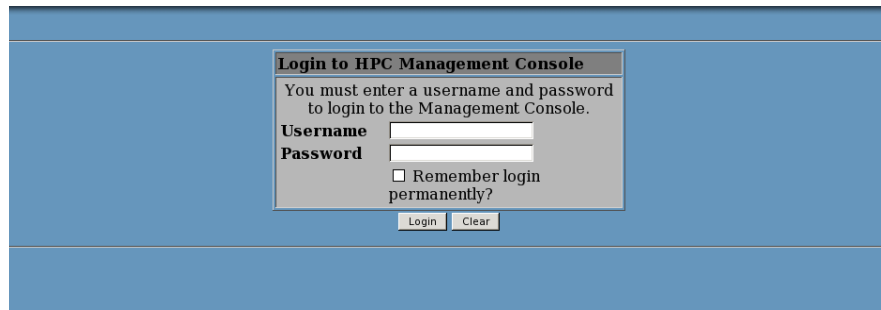
```
service sashpcmc status
```

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

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

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

The Login page appears.



Login to HPC Management Console

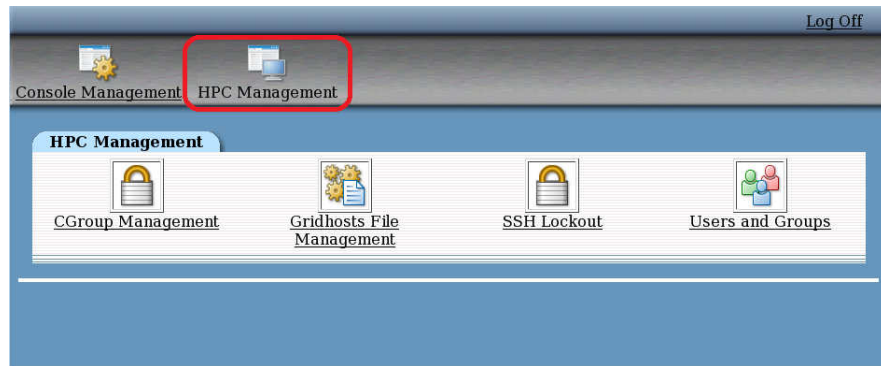
You must enter a username and password to login to the Management Console.

Username

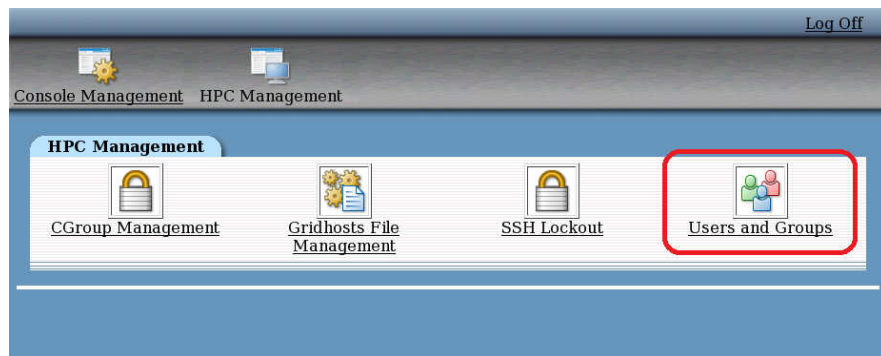
Password

☐ Remember login permanently?

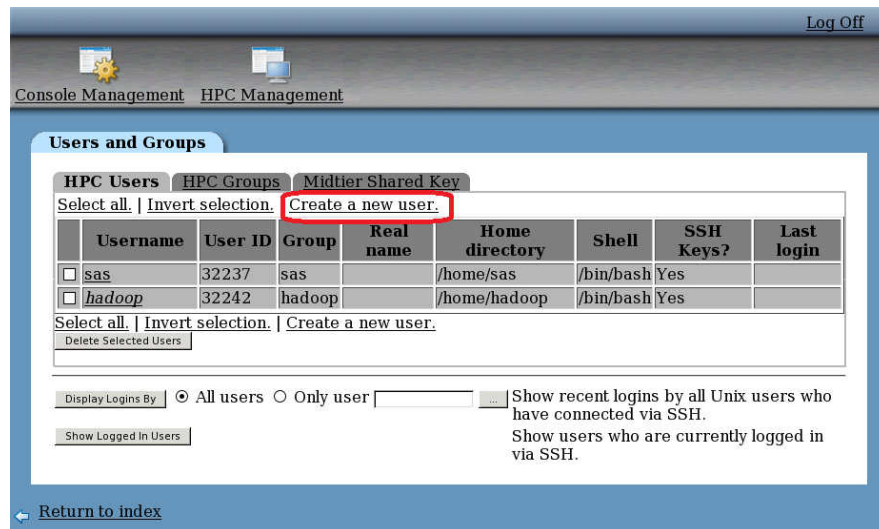
- 3 Log on to SAS High-Performance Computing Management Console.
The Console Management page appears.



- 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.

Create User

User Details

Username: sasdemo

User ID: ☒ Automatic ☐ Calculated

Real name:

Home directory: ☒ Automatic ☐ Directory

Shell: /bin/sh

Password: ☒ No password required ☐ Normal password

☐ Pre-encrypted password

Password Options

Password changed: Never Expiry date: / Jan /

Minimum days: Maximum days:

Warning days: Inactive days:

Group Membership

Primary group: ☒ New group with same name as user ☐ New group ☐ Existing group

Secondary groups: **All groups** cp-dns-ng mislSD misfin mismae mishr **In groups**

HPC Actions and Settings

Propagate User: ☒ Yes ☐ No

Generate and Propagate SSH Keys: ☒ Yes ☐ No

Add Shared Midtier Key: ☐ Yes ☒ No

Upon Creation..

Create home directory?: ☒ Yes ☐ No

Copy template files to home directory?: ☒ Yes ☐ No

Create

[Return to users and groups list](#)

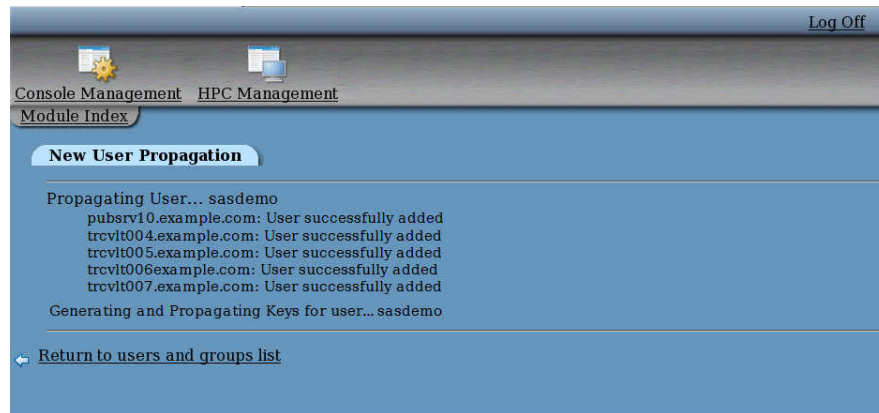
- 7 Enter information for the new user, using the security policies in place at your site.

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.



4

Modifying Co-Located Hadoop

<i>Infrastructure Deployment Process Overview</i>	37
<i>Overview of Modifying Co-located Hadoop</i>	38
<i>Modify the Existing Apache Hadoop Cluster, Version 2.7</i>	39
<i>Modify the Existing Cloudera Hadoop Cluster, Version 5</i>	41
<i>Modify the Existing Cloudera Hadoop Cluster, Version 4</i>	44
<i>Modify the Existing Hortonworks Data Platform Hadoop Cluster</i>	46
<i>Modify the Existing IBM BigInsights Hadoop Cluster</i>	48
<i>Modify the Existing MapR Hadoop Cluster</i>	50
<i>Modify the Existing Pivotal HD Hadoop Cluster</i>	50
<i>Hadoop Configuration Step for SAS Visual Analytics</i>	52

Infrastructure Deployment Process Overview

Modifying a co-located Hadoop cluster is an optional fifth of eight steps required to install and configure the SAS High-Performance Analytics infrastructure.

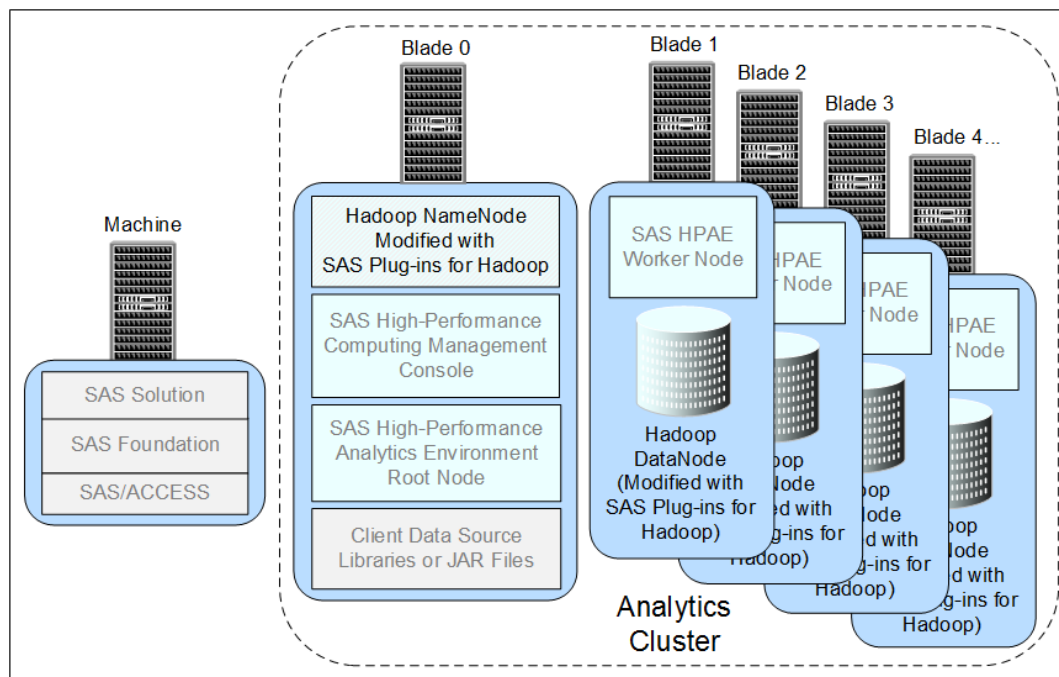
1. Create a SAS Software Depot.
2. Check for documentation updates.
3. Prepare your analytics cluster.
4. (Optional) Deploy SAS High-Performance Computing Management Console.
- **5. (Optional) Modify co-located Hadoop.**
6. Deploy the SAS High-Performance Analytics environment.
7. (Optional) Deploy the SAS Embedded Process for Hadoop.
8. (Optional) Configure the analytics environment for a remote parallel connection.

Overview of Modifying Co-located Hadoop

Some solutions, such as SAS Visual Analytics, rely on a SAS data store that is co-located with the SAS High-Performance Analytics environment on the analytics cluster.

The following figure shows the analytics environment co-located on a pre-existing Hadoop cluster. In this document, the NameNode is deployed on blade 0.

Figure 4.1 Analytics Cluster Co-located on the Hadoop Cluster



The SAS Plug-ins for Hadoop component provides services to your supported pre-existing Hadoop distribution that enable the SAS High-Performance Analytics environment to write SASHDAT file blocks evenly across the HDFS file system. This even distribution provides a balanced workload across the machines in the cluster and enables SAS analytic processes to read SASHDAT tables at very impressive rates.

With the exception of MapR Hadoop, modifying your supported Hadoop cluster for the analytics environment consists of the following steps:

- 1 Make sure that your Hadoop distribution meets the [requirements on page 21](#).
- 2 Copy the SAS Plug-ins for Hadoop package to a temporary location and untar it.
- 3 Propagate several files from the package to specific Hadoop directories on every machine in your cluster.
- 4 Restart the HDFS service and any dependencies.

The SAS High-Performance Analytics infrastructure supports the following Hadoop distributions:

- [Apache Hadoop, version 2.7 on page 39](#)
- [Cloudera Hadoop, version 4 on page 44](#)
- [Cloudera Hadoop, version 5 on page 41](#)
- [Hortonworks Data Platform Hadoop on page 46](#)
- [IBM BigInsights Hadoop on page 48](#)
- [MapR Hadoop on page 50](#)
- [Pivotal HD Hadoop on page 50](#)

For more information about SAS and Hadoop, see:

- [SAS 9.4 Supported Hadoop Distributions](#)
- [SAS 9.4 Support for Hadoop](#)
- [SAS and Hadoop Technology: Deployment Scenarios](#)
- [SAS and Hadoop Technology: Overview](#)

Modify the Existing Apache Hadoop Cluster, Version 2.7

If you want to co-locate the SAS High-Performance Analytics environment with a pre-existing Apache Hadoop cluster, you can modify your cluster with files from the SAS Plug-ins for Hadoop package. Apache Hadoop modified with SAS Plug-ins for Hadoop enables the SAS High-Performance Analytics environment to write SASHDAT file blocks evenly across the HDFS file system.

To modify line 2.7 or later of Apache Hadoop with SAS Plug-ins for Hadoop, follow these steps:

- 1 Make sure that you have reviewed all of the information contained in the section [“Requirements for Co-located Hadoop” on page 21](#).
- 2 Log on to the Hadoop NameNode machine (blade 0) as root.
- 3 Create a directory, `sas`, under `$HADOOP_HOME/share/hadoop`.
- 4 The software that is needed for SAS Plug-ins for Hadoop is available from within the SAS Software Depot that was created by your site’s depot administrator:

```
depot-installation-location/standalone_installs/  
SAS_Plug-ins_for_Hadoop/1_0/Linux_for_x64/hdatplugins.tar.gz
```

- 5 Copy the `hdatplugins.tar.gz` file to a temporary location and extract it:

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

A directory that is named `hdatplugins` is created.

- 6 Propagate the following three JAR files in `hdatplugins` to the `$HADOOP_HOME/share/hadoop/sas` directory on each machine in the Apache Hadoop cluster:

- sas.lasr.jar
- sas.lasr.hadoop.jar
- sas.grid.provider.yarn.jar

TIP If you have already installed the SAS High-Performance Computing Management Console or the SAS High-Performance Analytics environment, you can issue a single **simcp** command to propagate JAR files across all machines in the cluster. For example:

```
/opt/TKGrid/bin/simcp /tmp/hdatplugins/sas.lasr.jar
$HADOOP_HOME/share/hadoop/sas
/opt/TKGrid/bin/simcp /tmp/hdatplugins/sas.lasr.hadoop.jar
$HADOOP_HOME/share/hadoop/sas
/opt/TKGrid/bin/simcp /tmp/hdatplugins/sas.grid.provider.yarn.jar
$HADOOP_HOME/share/hadoop/sas
```

For more information, see [Appendix 2, “SAS High-Performance Analytics Infrastructure Command Reference,”](#) on page 85.

- 7 Propagate `saslasrfd` in `hdatplugins` to the `$HADOOP_HOME/bin` directory on each machine in the Apache Hadoop cluster.
- 8 Propagate `SAS_VERSION` in `hdatplugins` to the `$HADOOP_HOME/` directory on each machine in the Apache Hadoop cluster.
- 9 Define the following properties in `$HADOOP_HOME/etc/hadoop/hdfs-site.xml` and propagate the changes across all nodes in your Hadoop cluster:

```
<property>
  <name>dfs.namenode.plugins</name>
  <value>com.sas.lasr.hadoop.NameNodeService</value>
</property>
<property>
  <name>dfs.datanode.plugins</name>
  <value>com.sas.lasr.hadoop.DataNodeService</value>
</property>
<property>
  <name>com.sas.lasr.hadoop.fileinfo</name>
  <value>ls -l {0}</value>
  <description>The command used to get the user, group, and permission
  information for a file.
  </description>
</property>
<property>
  <name>com.sas.lasr.service.allow.put</name>
  <value>true</value>
  <description>Flag indicating whether the PUT command is enabled when
  running as a service. The default is false.
  </description>
</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>
```

Note: You can change the port for the SAS name node and data node plug-ins. This example shows the default ports (15452 and 15453, respectively).

- 10** Add the plug-in JAR files to the Hadoop class path in `$HADOOP_HOME/etc/hadoop/hadoop-env.sh` and propagate the changes across all nodes in your Hadoop cluster:

```
for f in $HADOOP_HOME/share/hadoop/sas/*.jar; do
  if [ "$HADOOP_CLASSPATH" ]; then
    export HADOOP_CLASSPATH=$HADOOP_CLASSPATH:$f
  else
    export HADOOP_CLASSPATH=$f
  fi
done
```

- 11** Restart the HDFS service and any dependencies.

- 12** If you are deploying SAS Visual Analytics, see [“Hadoop Configuration Step for SAS Visual Analytics”](#) on page 52.

Modify the Existing Cloudera Hadoop Cluster, Version 5

If you want to co-locate the SAS High-Performance Analytics environment with a pre-existing Cloudera 5 Hadoop (CDH) cluster, you can modify your cluster with files from the SAS Plug-ins for Hadoop package. CDH modified with SAS Plug-ins for Hadoop enables the SAS High-Performance Analytics environment to write SASHDAT file blocks evenly across the HDFS file system.

- 1** Untar the SAS Plug-ins for Hadoop tarball, and propagate five files (identified in the following steps) on every machine in your CDH cluster:

- a** Navigate to the SAS Plug-ins for Hadoop tarball in your SAS Software depot:

```
cd depot-installation-location/standalone_installs/
SAS_Plug-ins_for_Hadoop/1_0/Linux_for_x64/hdatplugins.tar.gz
```

- b** Copy `hdatplugins.tar.gz` to a temporary location where you have Write access.

- c** Untar `hdatplugins.tar.gz`:

```
tar xzf hdatplugins.tar.gz
```

The `hdatplugins` directory is created.

- d** Propagate the following three JAR files in `hdatplugins` into the CDH library path on every machine in the CDH cluster:

- `sas.lasr.jar`
- `sas.lasr.hadoop.jar`

- `sas.grid.provider.yarn.jar`

TIP If you have already installed the SAS High-Performance Computing Management Console or the SAS High-Performance Analytics environment, you can issue a single **simcp** command to propagate JAR files across all machines in the cluster. For example:

```
/opt/TKGrid/bin/simcp /tmp/hdatplugins/sas.lasr.jar
/opt/cloudera/parcels/CDH-5.0.0-0.cdh5b1.p0.57/lib/hadoop/lib/
/opt/TKGrid/bin/simcp /tmp/hdatplugins/sas.lasr.hadoop.jar
/opt/cloudera/parcels/CDH-5.0.0-0.cdh5b1.p0.57/lib/hadoop/lib/
/opt/TKGrid/bin/simcp /tmp/hdatplugins/sas.grid.provider.yarn.jar
/opt/cloudera/parcels/CDH-5.0.0-0.cdh5b1.p0.57/lib/hadoop/lib/
```

For more information, see [Appendix 2, “SAS High-Performance Analytics Infrastructure Command Reference,”](#) on page 85.

- e Propagate `saslasrfd` in `hdatplugins` into the CDH `bin` directory on every machine in the CDH cluster. For example:

```
/opt/TKGrid/bin/simcp saslasrfd
/opt/cloudera/parcels/CDH-5.0.0-0.cdh5b1.p0.57/lib/hadoop/bin/
```

- f Propagate `SAS_VERSION` in `hdatplugins` to the `$HADOOP_HOME` directory on each machine in the CDH cluster.

2 Log on to the Cloudera Manager as an administrator.

- 3 In `dfs.namenode.plugins`, add the following to the plug-in configuration for the NameNode:

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

- 4 In `dfs.datanode.plugins`, 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 Advanced Configuration Snippet (Safety Valve) 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** Add the following property to the **HDFS Client Advanced Configuration Snippet (Safety Valve) for `hdfs-site.xml`** under **Advanced** within the **Gateway Default Group**. Make sure that you change *path-to-data-dir* to the data directory location for your site (for example, `<value>file:///dfs/dn</value>`):

```
<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>
```

- 7** Add the location of `JAVA_HOME` to the **HDFS Client Environment Advanced Configuration Snippet for `hadoop-env.sh` (Safety Valve)**, located under **Advanced** in the **Gateway Default Group**. For example:

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

Note: When Cloudera Manager prioritizes the HDFS client configuration, the client safety valve is used. When Cloudera Manager prioritizes anything else (such as YARN), the service safety valve is used. Therefore, updating both safety values is the best practice. For more information, see [Cloudera documentation](#).

- 8** Save your changes and deploy the client configuration to each host in the cluster.
- 9** Restart the HDFS service and any dependencies in Cloudera Manager.
- 10** Create and mode the `/test` directory in HDFS for testing the cluster with SAS test jobs. You might need to set `HADOOP_HOME` first, and you must run the following commands as the user running HDFS (typically, `hdfs`).

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

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

- 11** If you are deploying SAS Visual Analytics, see “[Hadoop Configuration Step for SAS Visual Analytics](#)” on page 52.

Modify the Existing Cloudera Hadoop Cluster, Version 4

If you want to co-locate the SAS High-Performance Analytics environment with a pre-existing Cloudera 4 Hadoop (CDH) cluster, you can modify your cluster with files from the SAS Plug-ins for Hadoop package. CDH modified with SAS Plug-ins for Hadoop enables the SAS High-Performance Analytics environment to write SASHDAT file blocks evenly across the HDFS file system.

TIP In Cloudera 4.2 and earlier, you must install the enterprise license, even if you are below the stated limit of 50 nodes in the Hadoop cluster for requiring a license.

- 1 Untar the SAS Plug-ins for Hadoop tarball, and propagate five files (identified in the following steps) on every machine in your CDH cluster:

- a Navigate to the SAS Plug-ins for Hadoop tarball in your SAS Software depot:

```
cd depot-installation-location/standalone_installs/  
SAS_Plug-ins_for_Hadoop/1_0/Linux_for_x64/
```

- b Copy hdatplugins.tar.gz to a temporary location where you have Write access.

- c Untar hdatplugins.tar.gz:

```
tar xzf hdatplugins.tar.gz
```

The **hdatplugins** directory is created.

- d Propagate the following three JAR files in **hdatplugins** into the CDH library path on every machine in the CDH cluster:

- sas.lasr.jar
- sas.lasr.hadoop.jar
- sas.grid.provider.yarn.jar

TIP If you have already installed the SAS High-Performance Computing Management Console or the SAS High-Performance Analytics environment, you can issue a single **simcp** command to propagate JAR files across all machines in the cluster. For example:

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

For more information, see [Appendix 2, “SAS High-Performance Analytics Infrastructure Command Reference,”](#) on page 85.

- e Propagate `saslasrfd` in `hdatplugins` into the CDH `bin` directory to the CDH cluster on every machine in the CDH cluster. For example:

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

- 2 Propagate `SAS_VERSION` in `hdatplugins` to the `$HADOOP_HOME` directory on each machine in the CDH cluster.

- 3 Log on to the Cloudera Manager as an administrator.

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

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

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

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

- 6 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>
```

- 7 Restart all Cloudera Manager services.

- 8 Add the following to the **HDFS Client Configuration Safety Valve**:

```
<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>
```

- 9 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
```

- 10 Save your changes and deploy the client configuration to each host in the cluster.
- 11 If you are deploying SAS Visual Analytics, see [“Hadoop Configuration Step for SAS Visual Analytics” on page 52](#).

TIP Remember the value of HADOOP_HOME as the SAS High-Performance Analytics environment prompts for this during its install. By default, these are the values for Cloudera:

- Cloudera 4.5:

```
/opt/cloudera/parcels/CDH-4.5.0-1.cdh4.5.0.p0.30
```

- Cloudera 4.2 and earlier:

```
/opt/cloudra/parcels/CDH-4.2.0-1.cdh4.2.0.po.10/lib/Hadoop
```

Modify the Existing Hortonworks Data Platform Hadoop Cluster

If you want to co-locate the SAS High-Performance Analytics environment with a pre-existing Hortonworks Data Platform (HDP) cluster, you can modify your cluster with files from the SAS Plug-ins for Hadoop package. HDP modified with SAS Plug-ins for Hadoop enables the SAS High-Performance Analytics environment to write SASHDAT file blocks evenly across the HDFS file system.

- 1 Log on to Ambari as an administrator, and stop all HDP services.
- 2 Untar the SAS Plug-ins for Hadoop tarball, and propagate five files (identified in the following steps) on every machine in your Hortonworks Hadoop cluster:
 - a Navigate to the SAS Plug-ins for Hadoop tarball in your SAS Software depot:


```
cd depot-installation-location/standalone_installs/  
SAS_Plug-ins_for_Hadoop/1_0/Linux_for_x64/
```
 - b Copy `hdatplugins.tar.gz` to a temporary location where you have Write access.
 - c Untar `hdatplugins.tar.gz`:


```
tar xzf hdatplugins.tar.gz
```

The `hdatplugins` directory is created.
 - d Propagate the following three JAR files in `hdatplugins` into the HDP library path on every machine in the HDP cluster:
 - `sas.lasr.jar`

- `sas.lasr.hadoop.jar`
- `sas.grid.provider.yarn.jar`

TIP If you have already installed the SAS High-Performance Computing Management Console or the SAS High-Performance Analytics environment, you can issue a single **simcp** command to propagate JAR files across all machines in the cluster. For example:

```
/opt/TKGrid/bin/simcp /tmp/hdatplugins/sas.lasr.jar
/usr/hdp/2.2.0.0-2041/hadoop/lib
/opt/TKGrid/bin/simcp /tmp/hdatplugins/sas.lasr.hadoop.jar
/usr/hdp/2.2.0.0-2041/hadoop/lib
/opt/TKGrid/bin/simcp /tmp/hdatplugins/sas.grid.provider.yarn.jar
/usr/hdp/2.2.0.0-2041/hadoop/lib
```

For more information, see [Appendix 2, “SAS High-Performance Analytics Infrastructure Command Reference,”](#) on page 85.

- e Propagate `saslasrfd` in `hdatplugins` into the HDP `bin` directory on every machine in the HDP cluster. For example:

```
/opt/TKGrid/bin/simcp saslasrfd /usr/hdp/2.2.0.0-2041/hadoop/bin/
```

- 3 Propagate `SAS_VERSION` in `hdatplugins` to the `$HADOOP_HOME` directory on each machine in the HDP cluster.
- 4 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.service.allow.put

`true`

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

`15452`

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

`15453`

dfs.namenode.fs-limits.min-block-size

`0`

- 5 Save the properties and restart all HDP services.
- 6 If you are deploying SAS Visual Analytics, see [“Hadoop Configuration Step for SAS Visual Analytics”](#) on page 52.

Modify the Existing IBM BigInsights Hadoop Cluster

If you want to co-locate the SAS High-Performance Analytics environment with a pre-existing IBM BigInsights Hadoop cluster, you can modify your cluster with files from the SAS Plug-ins for Hadoop package. BigInsights modified with SAS Plug-ins for Hadoop enables the SAS High-Performance Analytics environment to write SASHDAT file blocks evenly across the HDFS file system.

- 1 Untar the SAS Plug-ins for Hadoop tarball, and propagate five files (identified in the following steps) on every machine in your BigInsights Hadoop cluster:

- a Navigate to the SAS Plug-ins for Hadoop tarball in your SAS Software depot:

```
cd depot-installation-location/standalone_installs/
SAS_Plug-ins_for_Hadoop/1_0/Linux_for_x64/
```

- b Copy hdatplugins.tar.gz to a temporary location where you have Write access.

- c Untar hdatplugins.tar.gz:

```
tar xzf hdatplugins.tar.gz
```

The `hdatplugins` directory is created.

- d Propagate the following three JAR files in `hdatplugins` into the library path on every machine in the BigInsights cluster:

- `sas.lasr.jar`
- `sas.lasr.hadoop.jar`
- `sas.grid.provider.yarn.jar`

Note: The default location for `HADOOP_HOME` is `/opt/ibm/biginsights/IHC`. The default location for `BIGINIGHTS_HOME` is `/opt/ibm/biginsights`.

TIP If you have already installed the SAS High-Performance Computing Management Console or the SAS High-Performance Analytics environment, you can issue a single **simcp** command to propagate JAR files across all machines in the cluster. For example:

```
/opt/TKGrid/bin/simcp /tmp/hdatplugins/sas.lasr.jar
$HADOOP_HOME/share/hadoop/hdfs/lib
/opt/TKGrid/bin/simcp /tmp/hdatplugins/sas.lasr.hadoop.jar
$HADOOP_HOME/share/hadoop/hdfs/lib
/opt/TKGrid/bin/simcp /tmp/hdatplugins/sas.grid.provider.yarn.jar
$HADOOP_HOME/share/hadoop/hdfs/lib
```

For more information, see [Appendix 2, “SAS High-Performance Analytics Infrastructure Command Reference,”](#) on page 85.

- e Propagate `saslasrfd` in `hdatplugins` into the `$HADOOP_HOME/bin` directory on every machine in the BigInsights cluster. For example:

```
/opt/TKGrid/bin/simcp saslasrfd $HADOOP_HOME/bin
```

- 2 Propagate `SAS_VERSION` in `hdatplugins` to the `$HADOOP_HOME` directory on each machine in the HDP cluster.
- 3 On the machine where you initially installed BigInsights, add the following properties for SAS for the HDFS configuration to the file `$BIGINSIGHTS_HOME/hdm/hadoop-conf-staging/hdfs-site.xml`. Adjust values appropriately for your deployment:

```
<property>
<name>dfs.namenode.plugins</name>
<value>com.sas.lasr.hadoop.NameNodeService</value>
</property>
<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>
```

- 4 Synchronize this new configuration by running the following command on the machine where you initially deployed BigInsights:

```
$BIGINSIGHTS_HOME/bin/syncconf.sh.
```

- 5 On the machine where you initially deployed BigInsights, log on as the `biadmin` user and run the following commands to restart the cluster with the new configuration:

```
stop-all.sh
```

```
start-all.sh
```

- 6 Note the location of `HADOOP_HOME`. You will need to refer to this value when installing the SAS High-Performance Analytics environment.
- 7 If you are deploying SAS Visual Analytics, see [“Hadoop Configuration Step for SAS Visual Analytics”](#) on page 52.

Modify the Existing MapR Hadoop Cluster

To configure your existing MapR Hadoop cluster to interoperate with the SAS High-Performance Analytics environment, follow these steps:

- 1 On all the SAS High-Performance Analytics environment nodes (the name node and all its worker nodes), create an identical mountpoint for your MapR file system.

For more information, see <http://doc.mapr.com/display/MapR/Setting+Up+MapR+NFS>.

- 2 When deploying the SAS High-Performance Analytics environment, enter this NFS mountpoint when prompted (for example, `/mapr/my.cluster.com`).

For more information, see [Table 5.2 on page 59](#).

- 3 If you are deploying SAS Visual Analytics, see “[Hadoop Configuration Step for SAS Visual Analytics](#)” on page 52.

After you have deployed SAS, for SAS LASR Server, add a server that loads from the NFS mountpoint. You can populate data in all the normal ways: copy to the SASHDAT engine, save an in-memory table, or use PROC IMXfer from a server running on a conventional Hadoop distribution.

For more information, see the [SAS LASR Analytic Server Reference Guide](#).

Modify the Existing Pivotal HD Hadoop Cluster

If you want to co-locate the SAS High-Performance Analytics environment with a pre-existing Pivotal HD (PHD) Hadoop cluster, you can modify your cluster with files from the SAS Plug-ins for Hadoop package. PHD modified with SAS Plug-ins for Hadoop enables the SAS High-Performance Analytics environment to write SASHDAT file blocks evenly across the HDFS file system.

- 1 Log on to Pivotal Command Center (PCC) as gpadmin. (The default password is Gpadmin1.)
- 2 Untar the SAS Plug-ins for Hadoop tarball, and propagate five files (identified in the following steps) on every machine in your PHD cluster:

- a Navigate to the SAS Plug-ins for Hadoop tarball in your SAS Software depot:

```
cd depot-installation-location/standalone_installs/
SAS_Plug-ins_for_Hadoop/1_0/Linux_for_x64/
```

- b Copy hdatplugins.tar.gz to a temporary location where you have Write access.

- c Untar hdatplugins.tar.gz:

```
tar xzf hdatplugins.tar.gz
```

The **hdatplugins** directory is created.

- d** Propagate the following three JAR files in **hdatplugins** into the library path on every machine in the PHD cluster:

- **sas.lasr.jar**
- **sas.lasr.hadoop.jar**
- **sas.grid.provider.yarn.jar**

TIP If you have already installed the SAS High-Performance Computing Management Console or the SAS High-Performance Analytics environment, you can issue a single **simcp** command to propagate JAR files across all machines in the cluster. For example:

```
/opt/TKGrid/bin/simcp /tmp/hdatplugins/sas.lasr.jar
/usr/lib/gphd/hadoop/lib/
/opt/TKGrid/bin/simcp /tmp/hdatplugins/sas.lasr.hadoop.jar
/usr/lib/gphd/hadoop/lib/
/opt/TKGrid/bin/simcp /tmp/hdatplugins/sas.grid.provider.yarn.jar
/usr/lib/gphd/hadoop/lib/
```

For more information, see [Appendix 2, “SAS High-Performance Analytics Infrastructure Command Reference,”](#) on page 85.

- e** Propagate **saslasrfd** in **hdatplugins** into the PHD **bin** directory on every machine in the PHD cluster. For example:

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

- 3** Propagate **SAS_VERSION** in **hdatplugins** to the **\$HADOOP_HOME** directory on each machine in the PHD cluster.
- 4** In the PCC, for YARN, make sure that Resource Manager, History Server, and Node Managers have unique host names.
- 5** In the PCC, make sure that the Zookeeper Server contains a unique host name.
- 6** Add the following properties for SAS for the HDFS configuration to the file **hdfs-site.xml**:

```
<property>
<name>dfs.namenode.plugins</name>
<value>com.sas.lasr.hadoop.NameNodeService</value>
</property>
<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>
```

- 7 Save your changes.
- 8 Restart your cluster using PCC and verify that HDFS is running in the dashboard.
- 9 If you are deploying SAS Visual Analytics, see [“Hadoop Configuration Step for SAS Visual Analytics” on page 52](#).

Hadoop Configuration Step for SAS Visual Analytics

If you deploying SAS Visual Analytics, then there are two directories that you must create in the HDFS file system, with the following permissions:

```
hadoop fs -mkdir /hps
hadoop fs -mkdir /vapublic
hadoop fs -chmod 777 /hps
hadoop fs -chmod 1777 /vapublic
```

5

Deploying the SAS High-Performance Analytics Environment

<i>Infrastructure Deployment Process Overview</i>	53
<i>Overview of Deploying the Analytics Environment</i>	54
<i>Encrypting SASHDAT Files</i>	57
<i>Install the Analytics Environment</i>	58
<i>Configuring the Analytics Environment for SASHDAT Encryption</i>	61
<i>Validating the Analytics Environment Deployment</i>	62
Overview of Validating	62
Use simsh to Validate	63
Use MPI to Validate	63
<i>Resource Management for the Analytics Environment</i>	64
Resource Settings File	64
Request Memory with TKMPI_INFO	65

Infrastructure Deployment Process Overview

Installing and configuring the SAS High-Performance Analytics environment is the sixth of eight 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) Modify co-located Hadoop.
- ▶ **6. Deploy the SAS High-Performance Analytics environment.**
7. (Optional) Deploy the SAS Embedded Process for Hadoop.
8. (Optional) Configure the analytics environment for a remote parallel connection.

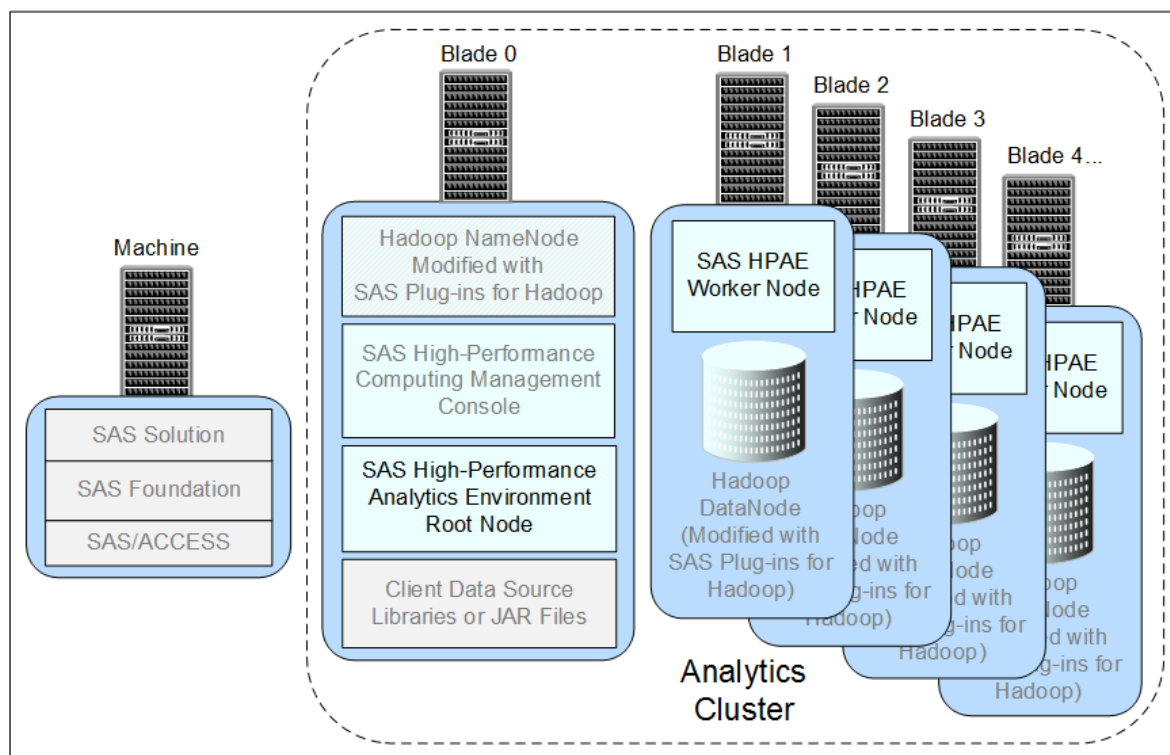
This chapter describes how to install and configure all of the components for the SAS High-Performance Analytics environment on the machines in the cluster.

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.

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

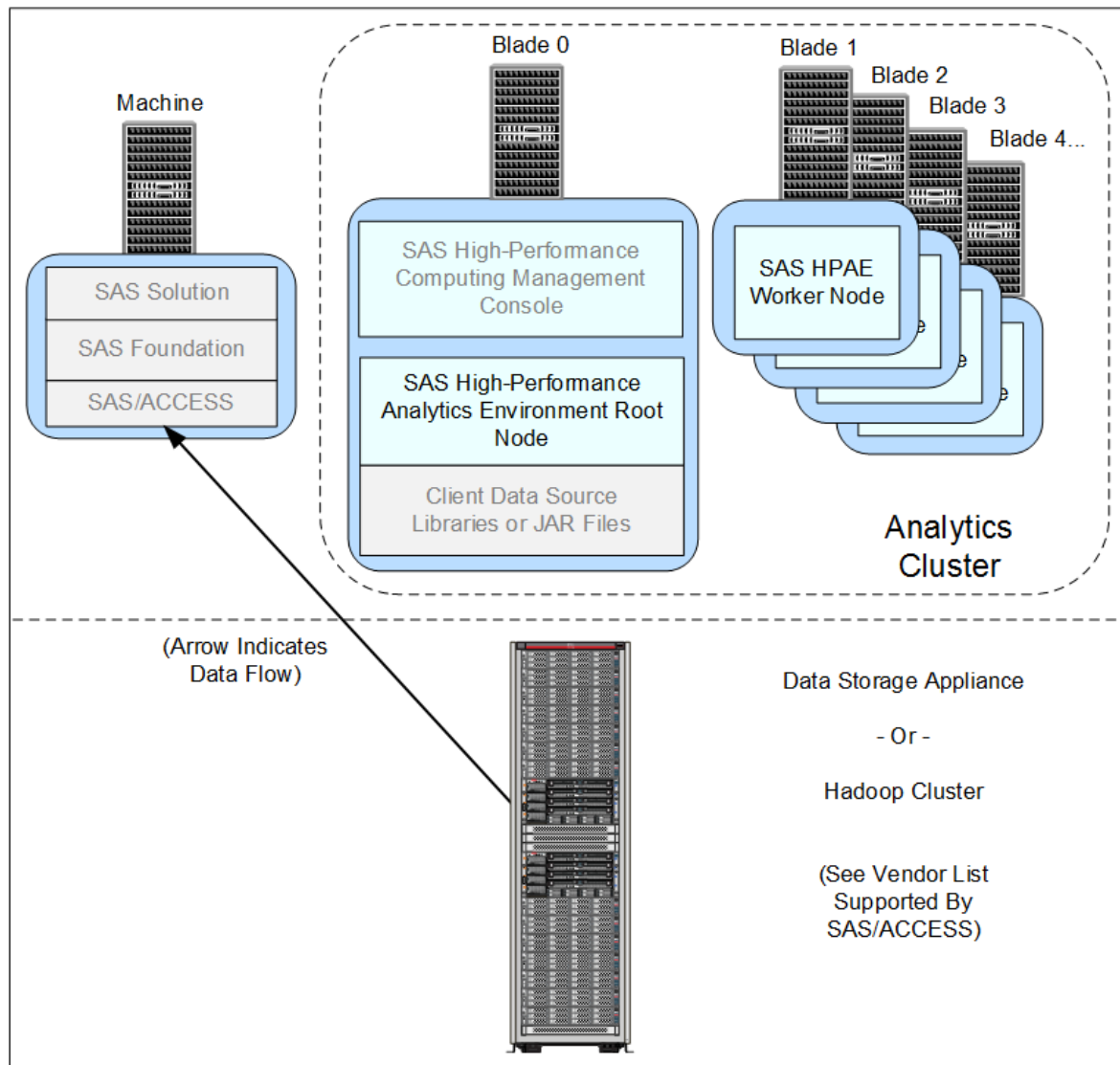
Figure 5.1 Analytics Environment Co-Located on the Hadoop Cluster



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:

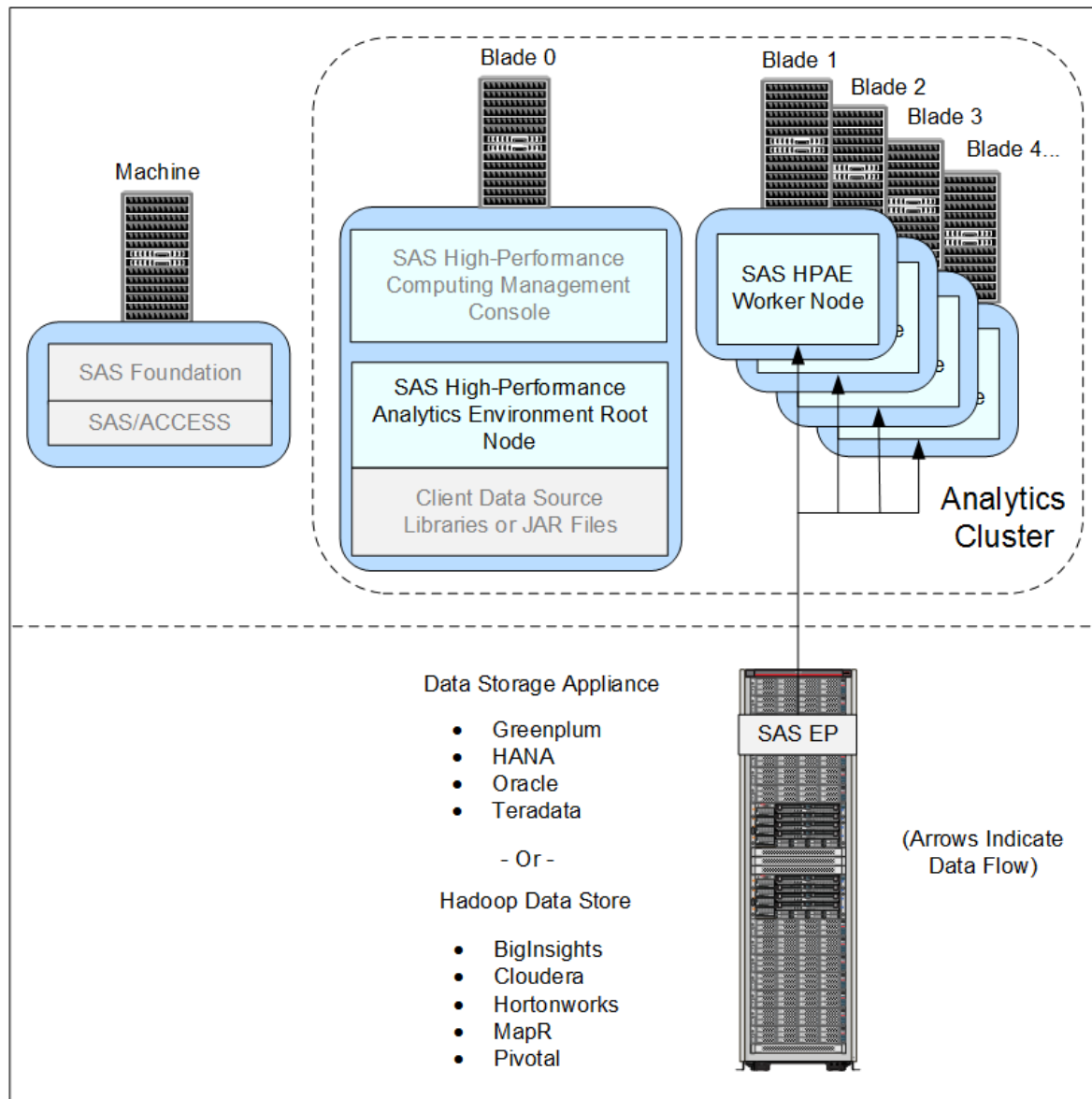
Figure 5.2 Analytics Environment Remote from Your Data Store (Serial Connection)



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 5.3 Analytics Environment Remote from Your Data Store (Parallel Connection)



The SAS High-Performance Analytics environment is packaged in separate executables. Refer to the following table for more information:

Table 5.1 *Installation and Configuration Packages for the SAS High-Performance Analytics Environment*

Order to install	Filename	Purpose
1	TKGrid_Linux_x86_64.sh	Analytics environment installation script for Red Hat Linux 6 and other equivalent, kernel-level Linux systems.
	TKGrid_Linux_x86_64_rhel5.sh	Analytics environment installation script for Red Hat Linux (pre-version 6) and SUSE Linux 10 systems.
2	TKTGDat.sh	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.
3 (optional)	TKGrid_SEC_x86_64.sh	Installation script for enabling the analytics environment to read and write encrypted SASHDAT files.
4 (optional)	TKGrid_REP_x86_64.sh	Script for configuring the SAS High-Performance Analytics environment with a SAS Embedded Process for Red Hat Linux 6 and other equivalent, kernel-level Linux systems.
	TKGrid_REP_x86_64_rhel5.sh	Script for configuring the SAS High-Performance Analytics environment with a SAS Embedded Process for Red Hat Linux (pre-version 6) and SUSE Linux 10 systems.

Encrypting SASHDAT Files

Starting with release 2.94, the SAS High-Performance Analytics environment supports reading and writing files using AES encryption with 256-bit keys. (This feature is very similar to the AES encryption provided by the SAS BASE Engine.) SASHDAT encryption is designed to bolster privacy protection for data at rest—that is, data stored in SASHDAT for analytic purposes.

Remember that SASHDAT data is typically not the system of record, but rather a copy of operational data that has been arranged for the purposes of analytics. In addition to encrypting data, many SAS users also anonymize their data when preparing it for analytics.

To enable the SAS High-Performance Analytics environment to read and write SASHDAT using encryption, you must install the TKGrid_SEC package. For more information, see [“Configuring the Analytics Environment for SASHDAT Encryption”](#) on page 61.

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 the SAS High-Performance Analytics Environment”](#) on page 21.
- 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/3_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 21.
- 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 5.2 Configuration Parameters for the TKGrid Shell Script

Parameter	Description
TKGrid Configuration Utility. Running on ' <i>machine-name</i> ' Using stdin for options. Shared install or replicate to each node? (Y=SHARED/n=replicated)	If you are installing to a local drive on each node, then specify <i>n</i> and press Enter 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 specify <i>y</i> and press Enter.
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 and press Enter. Otherwise, press Enter.
Enter NFS mount to MAPR directory (ie: /mapr/my.cluster.com, default is none).	<p>If you want the analytics environment to be able to read and write MapR data directly, enter the NFS mount here (for example, /mapr/my.cluster.com).</p> <p>The mount point must exist on all nodes, including the name node.</p> <p>The TKGrid script sets the environment variable, TKMPI_MAPRHDFS_PREFIX, to point to this share.</p> <p>For more information, see http://doc.mapr.com/display/MapR/Accessing+Data+with+NFS.</p>
Enter additional options to mpirun.	<p>If you have any mpirun options to add, specify them and press Enter.</p> <p>If you are using Kerberos, specify the following option and press Enter:</p> <pre>-genvlist `env sed -e s/=.*// sed / KRB5CCNAME/d tr -d '\n' `TKPATH,LD_LIBRARY_PATH</pre> <p>Note: Enter the above option on one line. Do not add any carriage returns or other whitespace characters.</p> <p>If you have no additional options, press Enter.</p>
Enter path to use for Utility files. (default is /tmp).	<p>SAS High-Performance Analytics applications might write scratch files. By default, these files are created in the /tmp directory. To accept the default, press Enter. Or, to redirect the files to a different location, specify the path and press Enter.</p> <p>Note: If the directory that you specified does not exist, you must create it manually.</p>
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) and press Enter. If your site does not use Hadoop, press Enter.
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, specify <i>y</i> and press Enter. Otherwise, specify <i>n</i> and press Enter.
Enter full path to machine list. The head node ' <i>head-node-machine-name</i> ' should be listed first.	Specify the name of the file that you created in the section " List the Machines in the Cluster or Appliance " (for example, /etc/gridhosts) and press Enter.

Parameter	Description
Enter maximum runtime for grid jobs (in seconds). Default 7200 (2 hours).	<p>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.</p> <p>To accept the default, press Enter. Or, specify a different maximum run time (in seconds) and press Enter.</p>
Enter value for UMASK. (default is unset.)	<p>To set <i>no</i> umask value, press Enter. Or, specify a umask value and press Enter.</p> <p>For more information, see “Consider Umask Settings” on page 22.</p>

- 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 TKGGrid 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 5.3 Configuration Prompts for the TKTGDat Shell Script

TKTG Configuration Utility. Running on 'machine-name' Using stdin for options. Shared install or replicate to each node? (Y=SHARED/n=replicated)	If you are installing to a local drive on each node, then specify n and press Enter 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 specify y and press Enter.
Enter full path to machine list.	Specify the name of the file that you created in the section “List the Machines in the Cluster or Appliance” (for example, /etc/gridhosts) and press Enter.

- 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)
```

If you want the installation program to perform the replication, specify **yes** and press Enter. If you are distributing the contents of the installation directory by some other technique, specify **no** and press Enter.

- 13** If you are planning to use the High-Performance Analytics environment in a locale other than English, you must copy the appropriate locale files from **SASFoundation/9.4/misc/tktg** to the **TKTGDat** directory on every machine in the analytics cluster.

In this example, the simultaneous command, **simcp**, is used to copy the Japanese locale files to the **TKTGDat** directory on each machine in the analytics cluster:

```
/opt/TKGrid/bin/simcp /opt/SASHome/SASFoundation/9.4/misc/tktg/jp* /opt/TKTGDat
```

- 14** Make one of the following choices:

- To enable the SAS High-Performance Analytics environment to read and write SASHDAT using encryption, proceed to [“Configuring the Analytics Environment for SASHDAT Encryption” on page 61](#).
- To configure the analytics environment for a SAS Embedded Process, proceed to [“Configuring for Access to a Data Store with a SAS Embedded Process” on page 75](#).
- To validate your analytics environment, proceed to [“Validating the Analytics Environment Deployment” on page 62](#).

Configuring the Analytics Environment for SASHDAT Encryption

In release 2.94, the SAS High-Performance Analytics environment supports reading and writing files using AES encryption with 256-bit keys. (This feature is very similar to the AES encryption provided by the SAS BASE Engine.)

Note: For U.S. export purposes, SAS designates each product based on the encryption algorithms and the product's functional capability. The ability to encrypt SASHDAT files is available to most commercial and government users inside and outside the U.S. However, some countries (for example, Russia, China, and France) have import restrictions on products that contain encryption, and the U.S. prohibits the export of encryption software to specific embargoed or restricted destinations.

To enable the SAS High-Performance Analytics environment to read and write SASHDAT using encryption, follow these steps:

- 1** 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_Encryption_Installation/3_5/Linux_for_x64*.
- 2** Copy **TKGrid_SEC_x86_64.sh** to the **/tmp** directory of the root node of the cluster.

- 3 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 21](#).

- 4 Change directories to the desired installation location, such as `/opt`.
- 5 Run the `TKGrid_SEC_x86_64` shell script in this directory.
- 6 Respond to the prompts from the shell script:

Table 5.4 Configuration Prompts for the `TKGrid_SEC_x86_64` 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 specify <code>n</code> and press Enter 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 specify <code>Y</code> and press Enter.
--	--

- 7 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.

Note: The contents of `TKGrid_SEC` must be distributed to every machine in the analytics cluster.

The shell script creates a `lib2` subdirectory and a file named `VERSION2`.

TIP If you are using Hadoop as your data provider, make sure that you follow the steps described for your distribution of Hadoop in [Chapter 4](#), [“Modifying Co-Located Hadoop,” on page 37](#).

- 8 To validate your analytics environment, proceed to [“Validating the Analytics Environment Deployment” on page 62](#).

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 63](#).

- “Use MPI to Validate” on page 63.

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 one of the machines in the analytics cluster.
- 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 [Chapter 6, “Configuring the Analytics Environment for a Remote Parallel Connection,”](#) on page 67.

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 [Chapter 6, “Configuring the Analytics Environment for a Remote Parallel Connection,”](#) on page 67.

Resource Management for the Analytics Environment

Resource Settings File

You can set limits on any TKGrid process running across the SAS High-Performance Analytics environment with a resource settings file supplied by SAS. Located in `/opt/TKGrid/`, `resource.settings` is in the format of a shell script. When the analytics environment starts, the environment variables contained in the file are set and last for the duration of the run.

Initially, all of the settings in `resource.settings` are commented. Uncomment the variables and add values that make sense for your site. For more information, see [“Using CGroups to Manage CPU” in SAS LASR Analytic Server: Reference Guide](#).

When you are finished editing, copy `resource.settings` to every machine in the analytics environment:

```
/opt/TKGrid/bin/simcp /opt/TKGrid/resource.settings /opt/TKGrid
```

If YARN is used on the cluster, then you can configure the analytics environment to participate in the resource accounting that YARN performs. For more information, see [“Managing Resources” in SAS LASR Analytic Server: Reference Guide](#).

`resource.settings` consists of the following:

```
# if [ "$USER" = "lasradm" ]; then
# Custom settings for any process running under the lasradm account.
#   export TKMPI_ULIMIT="-v 50000000"
#   export TKMPI_MEMSIZE=50000
#   export TKMPI_CGROUP="cgexec -g cpu:75"
# fi

# if [ "$TKMPI_APPNAME" = "lasr" ]; then
# Custom settings for a lasr process running under any account.
#   export TKMPI_ULIMIT="-v 50000000"
#   export TKMPI_MEMSIZE=50000
#   export TKMPI_CGROUP="cgexec -g cpu:75"

# Allow other users to read server and tables, but not add or term.
#   export TKMPI_UMASK=0033

# Allow no access by other users to lasr server.
#   export TKMPI_UMASK=0077

# To exclude from YARN resource manager.
#   unset TKMPI_RESOURCEMANAGER

# Use default nice for LASR
#   unset TKMPI_NICE
# fi
```



```
# if [ "$TKMPI_APPNAME" = "tklogis" ]; then
# Custom settings for a tklogis process running under any account.
#   export TKMPI_ULIMIT="-v 25000000"
#   export TKMPI_MEMSIZE=25000
#   export TKMPI_CGROUP="cgexec -g cpu:25"
#   export TKMPI_MAXRUNTIME=7200
# fi

# fi
# if [ "$TKMPI_INFO" = "LASRLOAD" ]; then
#   TKMPI_INFO is an environment variable that will be passed from
#   MVA SAS to the grid. It can be used to distinguish a
#   proc lasr create from a proc lasr add, by including
#   this line before the proc lasr add:
#   options set=TKMPI_INFO="LASRLOAD";
#   To exclude from YARN resource manager.
#   unset TKMPI_RESOURCEMANAGER
# fi
```

Request Memory with TKMPI_INFO

When programmers use TKMPI_INFO in their SAS code, the SAS High-Performance Analytics environment can better decide how much memory to request.

Consider this example: the \$TKMPI_APPNAME variable is set to `lasr` for both a SAS Analytic LASR Server (PROC LASR CREATE) and for a SAS Analytic LASR Server Proxy used when loading a table (PROC LASR ADD). This makes it impossible to set a YARN memory limit differently for these two cases. Most likely, a SAS Analytic LASR Server would want a large amount of memory and the proxy server would require a smaller amount.

Here is an example of how you might use TKMPI_INFO in a SAS program to solve the memory issue:

```
options set=TKMPI_INFO="LASRSTART";
proc lasr create port=17761;
performance nodes=2; run;

options set=TKMPI_INFO="LASRLOAD";
proc lasr add data=sashelp.cars port=17761; run
```

In `resource.settings`, you might add an entry similar to the following:

```
if [ "$TKMPI_APPNAME" = "lasr" ]; then
  if [ "$TKMPI_INFO" = "LASRSTART" ];
    export TKMPI_MEMSIZE=60000
  fi
  if [ "$TKMPI_INFO" = "LASRLOAD" ];
    export TKMPI_MEMSIZE=4000
  fi
fi
```

Note that TKMPI_INFO is not limited to SAS Analytic LASR Server. TKMPI_INFO can also be used for any other HPA PROC. You could use the variable to pass any type of information you need to `resource.settings` (for example, SMALL, MEDIUM, LARGE classes).

6

Configuring the Analytics Environment for a Remote Parallel Connection

Infrastructure Deployment Process Overview	67
Overview of Configuring the Analytics Environment for a Remote Parallel Connection	68
Preparing for a Remote Parallel Connection	69
Overview of Preparing for a Remote Parallel Connection	69
Prepare for Hadoop	70
Prepare for a Greenplum Data Computing Appliance	73
Prepare for a HANA Cluster	73
Prepare for an Oracle Exadata Appliance	74
Prepare for a Teradata Managed Server Cabinet	74
Configuring for Access to a Data Store with a SAS Embedded Process	75
Overview of Configuring for Access to a Data Store with a SAS Embedded Process	75
How the Configuration Script Works	75
Configure for Access to a Data Store with a SAS Embedded Process	76

Infrastructure Deployment Process Overview

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

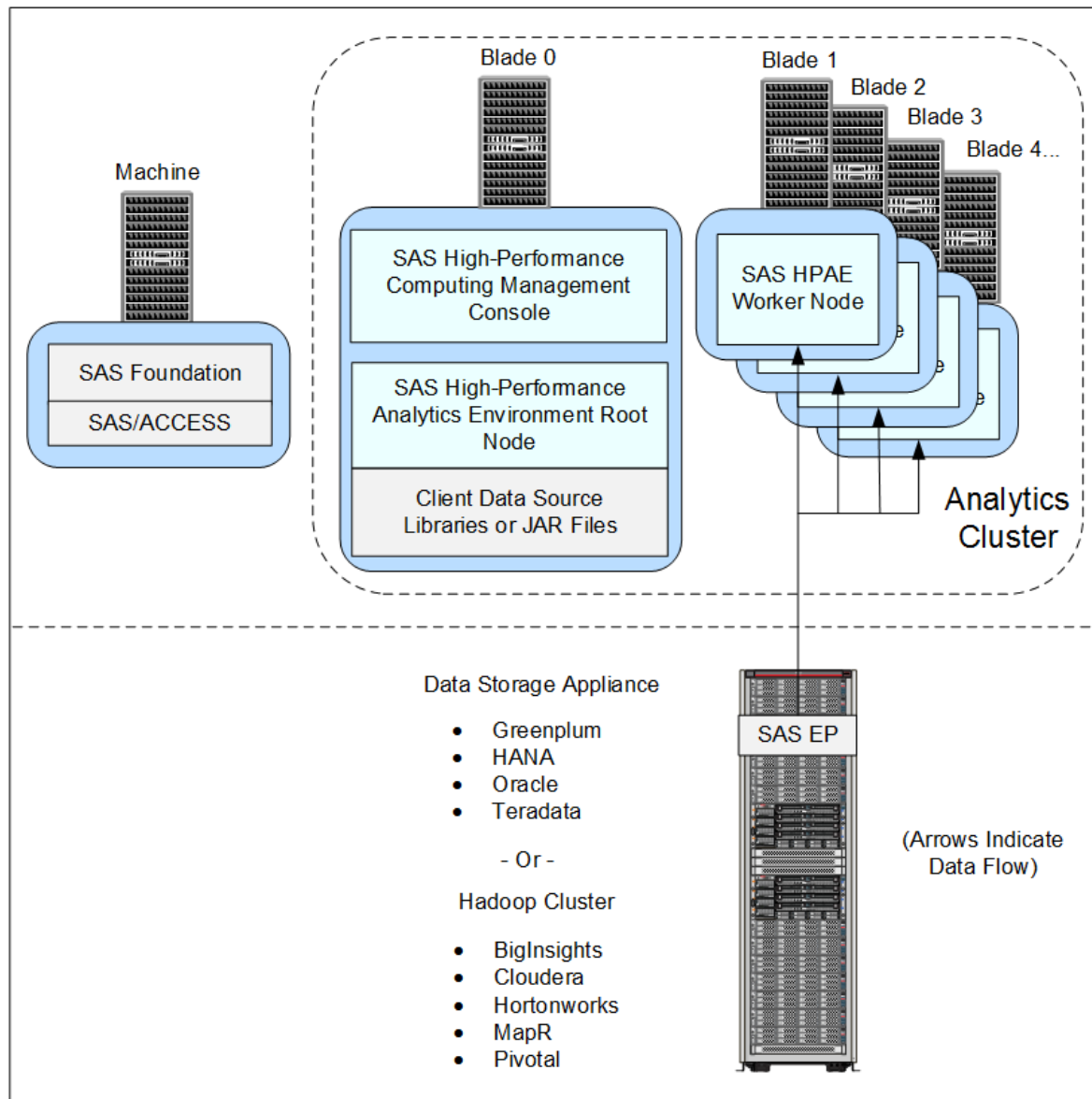
1. Create a SAS Software Depot.
2. Check for documentation updates.
3. Prepare your analytics cluster.
4. (Optional) Deploy SAS High-Performance Computing Management Console.
5. (Optional) Modify co-located Hadoop.
6. Deploy the SAS High-Performance Analytics environment.
7. (Optional) Deploy the SAS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

- **8. (Optional) Configure the analytics environment for a remote parallel connection.**

Overview of Configuring the Analytics Environment for a Remote Parallel Connection

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. After you have installed SAS/ACCESS and its embedded process, you configure the analytics environment for the particular access interface that you will use with a shell script, TKGrid_REP.

For information about installing the SAS Embedded Process, see the [SAS In-Database Products: Administrator's Guide](#).

Figure 6.1 Analytics Cluster Remote from Your Data Store (Parallel Connection)

Preparing for a Remote Parallel Connection

Overview of Preparing for a Remote Parallel Connection

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.

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

- 1 “Prepare for Hadoop” on page 70 .
- 2 “Prepare for a Greenplum Data Computing Appliance” on page 73.
- 3 “Prepare for a HANA Cluster” on page 73.
- 4 “Prepare for an Oracle Exadata Appliance” on page 74.
- 5 “Prepare for a Teradata Managed Server Cabinet” on page 74.

Prepare for Hadoop

Overview of Preparing for Hadoop

Note: In the third maintenance release of SAS 9.4, SAS Embedded Process supports the Cloudera, Hortonworks, IBM BigInsights, MapR, and Pivotal HD distributions of Hadoop. For more specific version information, see the [SAS 9.4 Support for Hadoop](#).

Preparing for a remote parallel connection to Hadoop consists of the following steps:

- 1 downloading and running a tracer script, `hadooptracer.py`, supplied by SAS.
`hadooptracer.py` traces the system calls of various Hadoop client tools and uses this data to identify the appropriate client JAR files that the SAS High-Performance Analytics environment requires for Hadoop client machine to Hadoop server environment connectivity.
 The script also copies these Hadoop client JAR files to a location on the Hadoop Hive node that you specify. You then manually copy this directory to the machine where you will deploy the SAS High-Performance environment root node. See [“Run the Hadoop Tracer Script” on page 71](#).
- 2 recording information about your Hadoop deployment that you will need when configuring the analytics environment for a remote data store.
 See [“Hadoop Checklists” on page 72](#).
- 3 determining the version of the JRE used by Hadoop and installing that same JRE on the analytics cluster.
 See [“Install the JRE on the Analytics Cluster” on page 73](#).

Prerequisites for the Hadoop Tracer Script

In order to run the [Hadoop tracer script](#), you must meet these prerequisites:

- Python 2.6 (or later) and `strace` must be installed.
- We recommend that your Hadoop Hive node machine must have a temporary directory named `tmp` under the root directory (`/tmp`).
 By default, the script uses a single directory (`/tmp`) as its temporary and output directories. However, you can change these using various script options.
- The user running the script must have the following:
 - a Linux account with SSH access (password or private key), to the Hive node or NameNode.

- authorization to issue HDFS and Hive commands.

Before the script is executed, a simple validation test is run to see whether HDFS (`hadoop`) and Hive (`hive`) commands can be issued. If the validation test fails, the script will not execute.

- a ticket (Kerberos only).

Run the Hadoop Tracer Script

To download and run the [Hadoop tracer script](#), follow these steps:

- 1 Make sure that your system meets the prerequisites.

See “[Prerequisites for the Hadoop Tracer Script](#)”.

- 2 On a machine from which you can also access your Hadoop Hive node, create a temporary directory where you will download the script.

Here is an example:

```
mkdir hadoopfiles_temp
```

- 3 Download the `hadooptracer.zip` file from the following FTP site to the directory that you created: <ftp://ftp.sas.com/techsup/download/blind/access/hadooptracer.zip>.

- 4 If there is not a `/tmp` directory on your Hadoop Hive node, create it.

By default, the script uses a single directory (`/tmp`) as its temporary and output directories. However, you can change these using various script options. See [Step 8](#).

- 5 Using a transfer method, such as PSFTP, SFTP, SCP, or FTP, transfer the ZIP file to the Hive node on your Hadoop cluster.

Here is an example:

```
scp /opt/sas/hadoopfiles_temp/hadooptracer.zip
root@hive_node.example.com:/tmp
```

- 6 On the Hadoop Hive node, change to the `/tmp` directory and unzip `hadooptracer.zip`.

Here is an example:

```
cd /tmp
tar xzf hadooptracer.zip
```

- 7 Enter the following command to grant Execute permissions on the script file:

```
chmod 755 ./hadooptracer.py
```

- 8 Run the script with these options:

```
python hadooptracer.py --filterby=latest --postprocess
```

Note: `hadooptracer.py` ignores `--postprocess` on Cloudera Hadoop clusters.

TIP For more information about script options, run the script with the `-h` option.

The script does the following:

- a** traces the system calls of various Hadoop client tools and uses this data to identify the appropriate client JAR files.
- b** copies the client JAR files to `/tmp/jars`.
- c** writes its log to `/tmp/hadooptracer.log`.

Note: Some error messages in the console output for `hadooptracer.py` are normal and do not necessarily indicate a problem with the JAR and configuration file collection process. However, if the files are not collected as expected or if you experience problems connecting to Hadoop with the collected files, contact SAS Technical Support and include the `hadooptracer.log` file.

- 9** When the script finishes executing, delete the following files:
 - `derby*.jar`
 - `spark-examples*.jar`
 - `ranger-plugins-audit*.jar`
 - `avatica*.jar`
 - `hadoop-0.20.2-dev-core*.jar`
- 10** Copy the JAR files that `hadooptracer.py` writes in `/tmp/jars` to a directory on the SAS High-Performance Analytics environment root node machine. As the analytics environment configuration script **prompts** you for this directory later, be sure to note it in [Table 6.1](#).

Hadoop Checklists

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.

Note: In the third maintenance release of SAS 9.4, the SAS Embedded Process supports the IBM BigInsights, Cloudera, Hortonworks, MapR, and Pivotal HD distributions of Hadoop. For more detailed information, see the [SAS Foundation system requirements documentation for your operating environment](#).

- 1** Record the path to the Hadoop client JAR files required by the analytics environment in the table that follows:

Table 6.1 Record the Location of the Hadoop Client JAR Files

Example	Actual Path of the Required Hadoop JAR Files on Your Analytics Environment Root Node
<code>/opt/hadoop_jars</code> (Hadoop client JAR files)	

Note: The location of the common and core JAR files listed in [Table 6.1](#) should be the same location that you copied the client JAR files to in [Step 10 on page 72](#).

- 2 Record the location (JAVA_HOME) of the 64-bit Java Runtime Engine (JRE) used by the analytics environment in the table that follows:

Table 6.2 Record the Location of the JRE Used by the SAS High-Performance Analytics Environment

Example	Actual Path of the JRE on Your Analytics Environment Root Node
<code>/opt/java/jre1.7.0_07</code>	

Note: This is the location of the JRE that the SAS High-Performance Analytics environment uses, not the location of the JRE that Hadoop uses.

Install the JRE on the Analytics Cluster

The SAS High-Performance Analytics environment requires a 64-bit Java Runtime Engine (JRE) when the environment is configured for a remote parallel connection.

Note: The JRE used by the analytics environment must match the version of the JRE used by your Hadoop cluster.

As the analytics environment configuration script [prompts](#) you for the location of this JRE (JAVA_HOME), be sure to note it in [Table 6.2](#).

Prepare for a Greenplum Data Computing 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 Computing 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 6.3 Record the Location of the Greenplum Client

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

Prepare for a HANA Cluster

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

- 1 Install the HANA client on blade 0 in your analytics cluster.

For more information, refer to your HANA documentation.

- 2 Record the path to the HANA client in the table that follows:

Table 6.4 Record the Location of the HANA Client

Example	Actual Path of the HANA Client on Your System
<hr/>	
	/usr/local/lib/hdbclient

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 6.5 Record the Location of the Oracle Client

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

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

Table 6.6 Record the Value of the Oracle TNS_ADMIN Environment Variable

Example	Oracle TNS_ADMIN Environment Variable Value on Your System
<hr/>	
	/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 6.7 Record the Location of the Teradata Client

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

Configuring for Access to a Data Store with a SAS Embedded Process

Overview of Configuring for Access to a Data Store with a SAS Embedded Process

The process involved for configuring the SAS High-Performance Analytics environment with a SAS Embedded Process consists of the following steps:

- 1 Prepare for the data provider that the analytics environment will query.
For more information, see [“Preparing for a Remote Parallel Connection” on page 69](#).
Note: Other third-party data providers besides Hadoop are supported. For more information, see the [SAS In-Database Products: Administrator’s Guide](#).
- 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 75](#).
- 3 Configure the analytics environment for a remote data store.
For more information, see [“Configure for Access to a Data Store with a SAS Embedded Process” on page 76](#).

How the Configuration Script Works

You configure the SAS High-Performance Analytics environment with a SAS Embedded Process 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 Access to a Data Store with a SAS Embedded Process

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 for a Remote Parallel Connection” 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 75](#).
- 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/3_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 21.

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 6.8 Configuration Parameters for the `TKGrid_REP` Shell Script

Parameter	Description
TKGrid Remote EP Addon Configuration Utility. Running on 'machine-name' Using stdin for options. Do you want to configure remote access to Teradata? (yes/NO)	If you are using a Teradata Managed Cabinet for your data provider, specify <code>yes</code> and press Enter. Otherwise, specify <code>no</code> and press Enter.
Enter path of Teradata client install. i.e.: <code>/opt/teradata/client/13.10</code>	If you specified <code>no</code> in the previous step, specify the path where the Teradata client was installed and press Enter. (This path was recorded earlier in Table 6.7 on page 75.)
Do you want to configure remote access to Greenplum? (yes/NO)	If you are using a Greenplum Data Computing Appliance for your data provider, specify <code>yes</code> and press Enter. Otherwise, specify <code>no</code> and press Enter.
Enter path of Greenplum client install. i.e.: <code>/usr/local/greenplum-db</code>	If you specified <code>no</code> in the previous step, specify the path where the Greenplum client was installed and press Enter. (This path was recorded earlier in Table 6.3 on page 73.)
Do you want to configure remote access to Hadoop? (yes/NO)	If you are using a Hadoop machine cluster for your data provider, specify <code>yes</code> and press Enter. Otherwise, specify <code>no</code> and press Enter.
Enter path of 64 bit JRE i.e.: <code>/usr/java/jdk1.7.0_09/jre</code>	If you chose <code>yes</code> in the previous step, specify the path where the JRE resides that analytics environment uses and press Enter. (This path was recorded earlier in Table 6.2 on page 73.)
Enter path of the directory (or directories separated by :) containing the Hadoop client JAR files.	Specify the path where the client Hadoop JAR files required by SAS reside and press Enter. (This path was recorded earlier in Table 6.1 on page 72.)
Enter any JRE Options you need added for the Java invocation, or just Enter if none.	If you need to add any JRE options, do so here (for example, - <code>Djava.security.auth.login.config=/opt/mapr/conf/mapr.login.conf</code> - <code>Djava.library.path=/opt/mapr/lib</code>).
Do you want to configure remote access to Oracle? (yes/NO)	If you are using an ORACLE Exadata appliance for your data provider, specify <code>yes</code> and press Enter. Otherwise, specify <code>no</code> and press Enter.

Parameter	Description
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 and press Enter. (This path was recorded earlier in Table 6.5 on page 74.)
Enter path of TNS_ADMIN, or just enter if not needed.	Enter the value of the Oracle TNS_ADMIN environment variable and press Enter. (This value was recorded earlier in Table 6.6 on page 74.)
Do you want to configure remote access to SAP HANA? (yes/NO)	If you are using a HANA cluster for your data provider, specify yes and press Enter. Otherwise, specify no and press Enter.
Enter path of HANA client install. i.e.: /usr/local/lib/hdbclient	Enter the path where the HANA client libraries reside and press Enter. (This path was recorded earlier in Table 6.4 on page 74.)
Shared install or replicate to each node? (Y=SHARED/n=replicated)	If you are installing to a local drive on each node, then select no and press Enter 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 specify yes and press Enter.
Enter path to TKGrid install	Specify the absolute path to where the analytics environment is installed and press Enter. 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 58.
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 analytics environment, specify the paths here and press Enter. Separate paths with a colon (:). If you have no paths to specify, press Enter.

- 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.](#)

- 11** To validate your analytics environment, proceed to “[Validating the Analytics Environment Deployment](#)” on page 62.

Appendix 1

Updating the SAS High-Performance Analytics Infrastructure

<i>Analytics Infrastructure Dependencies</i>	79
<i>Updating Co-located Hadoop</i>	79
Overview of Updating Co-located Hadoop	79
Updating Co-located Distributions of Hadoop	80
Updating SAS High-Performance Deployment for Hadoop (Prior to Version 3.0)	80
<i>Update the Analytics Environment</i>	83
<i>Updating the SAS High-Performance Computing Management Console</i>	84
Overview of Updating the Management Console	84
Update the Management Console Using RPM	84

Analytics Infrastructure Dependencies

Note the following dependencies when updating the SAS High-Performance Analytics infrastructure:

- If you update the analytics environment, you must also update your co-located Hadoop cluster.
- Update Hadoop first, followed by the analytics environment.

For more information, see [“Overview of Updating Co-located Hadoop”](#) on page 79.

Updating Co-located Hadoop

Overview of Updating Co-located Hadoop

Follow the instructions appropriate for your co-located Hadoop distribution:

- [“Updating SAS High-Performance Deployment for Hadoop \(Prior to Version 3.0\)”](#).
- [“Updating Co-located Distributions of Hadoop”](#).

Updating Co-located Distributions of Hadoop

This topic affects the following supported Hadoop distributions that are co-located with the SAS High-Performance Analytics environment:

- Apache Hadoop (version 2.7 and later)
- Cloudera Hadoop
- Hortonworks Data Platform Hadoop
- IBM BigInsights Hadoop
- MapR Hadoop
- Pivotal HD Hadoop

Note: If you are using SAS High-Performance Deployment for Hadoop, see [“Updating SAS High-Performance Deployment for Hadoop \(Prior to Version 3.0\)”](#).

Before upgrading your Hadoop distribution, remove or disable these properties to prevent HDFS startup failures:

- `dfs.namenode.plugins=com.sas.lasr.hadoop.NameNodeService`
- `dfs.datanode.plugins=com.sas.lasr.hadoop.DataNodeService`

After you have upgraded your co-located Hadoop, redo the steps in [Chapter 4, “Modifying Co-Located Hadoop,”](#) on page 37 appropriate for your Hadoop distribution. Performing these steps ensure that the SAS executable file and SAS JAR files are placed into the correct upgraded Hadoop paths and the configuration properties are in place to enable them.

Updating SAS High-Performance Deployment for Hadoop (Prior to Version 3.0)

Which Version of SAS High-Performance Deployment for Hadoop Have I Deployed?

SAS has discontinued SAS High-Performance Deployment for Hadoop.

If you are running a version of SAS High-Performance Deployment for Hadoop prior to version 3.0, then you must update your Hadoop distribution with files from the SAS Plug-ins for Hadoop package.

To verify the version of SAS High-Performance Deployment for Hadoop that you are running, run the following command:

```
more $HADOOP_HOME/SAS_VERSION
```

If the output matches the following, then you do not have to upgrade:

```
SAS Hadoop Extensions Version 3.0
```

Overview of Updating SAS High-Performance Deployment for Hadoop

Updating SAS High-Performance Deployment for Hadoop prior to version 3.0 consists of the following steps:

- 1 Follow the steps in “[Preparing to Update SAS High-Performance Deployment for Hadoop](#)”.
- 2 Follow the steps in “[Update SAS High-Performance Deployment for Hadoop](#)”, which are as follows:
 - 1 Copy the SAS Plug-ins for Hadoop package to a temporary location and untar it.
 - 2 Propagate several files from the package to specific Hadoop directories on every machine in your cluster.
 - 3 Restart the HDFS service and any dependencies.

Preparing to Update SAS High-Performance Deployment for Hadoop

Before starting the SAS High-Performance Deployment for Hadoop update, perform the following steps:

- 1 If one does not already exist, create a SAS Software Depot that contains the installation software that you will use to update Hadoop.
For more information, see [Creating a SAS Software Depot](#) in the *SAS Intelligence Platform: Installation and Configuration Guide*.
- 2 Log on to the Hadoop NameNode as the `hdfs` user.
- 3 Run the following command to make sure that the Hadoop file system is healthy: `hadoop fsck /`
Correct any issues before proceeding.
- 4 Stop any other processes, such as YARN, running on the Hadoop cluster.
Confirm that all processes have stopped across all the cluster machines. (You might have to become another user to have the necessarily privileges to stop all processes.)
- 5 As the `hdfs` user, run the command `$HADOOP_HOME/sbin/stop-dfs.sh` to stop HDFS daemons, and confirm that all processes have ceased on all the machines in the cluster.

TIP 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 **hdfs** user account, terminate them. 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 hdfs`.

- 6 Back up the Hadoop name directory (`hadoop-name` by default).
Perform a file system backup using tar (or whatever tool or process that your site uses for backups).

Update SAS High-Performance Deployment for Hadoop

To update versions prior to 3.0 of SAS High-Performance Deployment for Hadoop, follow these steps:

- 1 If your version of SAS High-Performance Deployment for Hadoop is version 3.0, then you do not have to update Hadoop. For more information, see [“Which Version of SAS High-Performance Deployment for Hadoop Have I Deployed?”](#) on page 80.
- 2 Make sure that you have completed all the steps in, [“Preparing to Update SAS High-Performance Deployment for Hadoop”](#) on page 81.

- 3 Log on to the Hadoop NameNode machine (blade 0) as root.

The software that is needed for SAS Plug-ins for Hadoop is available from within the SAS Software Depot that was created by your site’s depot administrator:

```
depot-installation-location/standalone_installs/  
SAS_Plug-ins_for_Hadoop/1_0/Linux_for_x64/hdatplugins.tar.gz
```

- 4 Copy the hdatplugins.tar.gz file to a temporary location and extract it:

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

A directory that is named `hdatplugins` is created.

- 5 Propagate the following three JAR files in `hdatplugins` to the `$HADOOP_HOME/share/hadoop/sas` directory on each machine in the Hadoop cluster:

- `sas.lasr.jar`
- `sas.lasr.hadoop.jar`
- `sas.grid.provider.yarn.jar`

CAUTION! If your distribution of SAS High-Performance Deployment for Hadoop is based on Apache Hadoop version 0.23, then you must propagate these JAR files to `$HADOOP_HOME/share/hadoop/hdfs/lib` instead.

TIP If you have already installed the SAS High-Performance Computing Management Console or the SAS High-Performance Analytics environment, you can issue a single **simcp** command to propagate JAR files across all machines in the cluster. For example:

```
/opt/TKGrid/bin/simcp /tmp/hdatplugins/sas.lasr.jar  
$HADOOP_HOME/share/hadoop/sas  
/opt/TKGrid/bin/simcp /tmp/hdatplugins/sas.lasr.hadoop.jar  
$HADOOP_HOME/share/hadoop/sas  
/opt/TKGrid/bin/simcp /tmp/hdatplugins/sas.grid.provider.yarn.jar  
$HADOOP_HOME/share/hadoop/sas
```

For more information, see [Appendix 2, “SAS High-Performance Analytics Infrastructure Command Reference,”](#) on page 85.

- 6 Propagate `saslasrfd` in `hdatplugins` to the `$HADOOP_HOME/share/hadoop/sas/bin` directory on each machine in the Hadoop cluster.

- 7 Propagate `SAS_VERSION` in `hdatplugins` to the `$HADOOP_HOME/share/hadoop/sas` directory on each machine in the Hadoop cluster.
- 8 Restart Hadoop by entering the following command:
`$HADOOP_HOME/sbin/start-dfs.sh`
- 9 If you are deploying SAS Visual Analytics, see [“Hadoop Configuration Step for SAS Visual Analytics” on page 52.](#)

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 58.](#)

- 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 58.](#)

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 the [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 If you are also updating your co-located Hadoop deployment, update Hadoop first, then the SAS High-Performance Analytics environment.

For more information, see [“Updating Co-located Hadoop” on page 79.](#)

- 2 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`.

- 3 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`.

- 4 Re-install the analytics environment using the shell script as described in “Install the Analytics Environment” on page 58.

Updating the SAS High-Performance Computing Management Console

Overview of Updating the Management Console

Starting in version 2.6 of SAS High-Performance Computing Management Console, there is no longer support for memory management through CGroups.

Before upgrading the management console to version 2.6 (or later), make sure that you manually record any memory settings and then clear them on the **CGroup Resource Management** page. You can manually transfer these memory settings to the SAS High-Performance Analytics environment resource settings file. Or, if you are implementing YARN, transfer these settings to YARN. For more information, see the [SAS LASR Analytic Server Reference Guide](#).

Update the Management Console Using RPM

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

- 1 Make sure that you have manually recorded and then cleared any memory settings in the management console. For more information, see “[Overview of Updating the Management Console](#)” on page 84.

- 2 Stop the server by entering the following command as the **root** user:

```
service sashpcmc stop
```

- 3 Update the management console using the following RPM command:

```
rpm -U --prefix=install-directory
/SAS-Software-Depot-root-directory/standalone_installs/
SAS_High-Performance_Computing_Management_Console/2_7/Linux_for_x64/
sashpcmc-2.7.x86_64.rpm
```

In this command, *install-directory* is the location where the management console is installed and *SAS-Software-Depot-root-directory* is the location where your SAS Software Depot resides.

- 4 Log on to the console to validate your update.

Appendix 2

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 the [nSAS High-Performance Computing Management Console: User's Guide](#).

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 3

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 4

Deploying on SELinux and IPTables

<i>Overview of Deploying on SELinux and IPTables</i>	89
<i>Prepare the Management Console</i>	90
SELinux Modifications for the Management Console	90
IPTables Modifications for the Management Console	90
<i>Prepare the Analytics Environment</i>	90
SELinux Modifications for the Analytics Environment	90
IPTables Modifications for the Analytics Environment	90
<i>Analytics Environment Post-Installation Modifications</i>	91
<i>iptables File</i>	91

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 Red Hat 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 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.

Here 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
```

Here 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.

Here 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
```

Appendix 5

Setting Up Passwordless Secure Shell (SSH)

What Is Passwordless SSH?

Secure Shell (SSH) is a network protocol that allows data to be exchanged using a secure channel between two networked devices. Passwordless SSH enables an identity to connect from one device to another without specifying a password. The identity can log on without a credential challenge, or it can invoke commands on the other device without a credential challenge.

Who Needs Passwordless SSH?

For a non-distributed server, passwordless SSH is not applicable.

For a distributed server, the requirements for passwordless SSH are as follows:

- Each user that needs to start and stop servers and load and unload tables must have an account that is configured for passwordless SSH on each machine in the cluster.
- If you use automated loading, the service account under which the scheduled task runs must be configured for passwordless SSH on each machine in the cluster. This is necessary to perform tasks such as starting and stopping the server and loading and unloading tables.
- For deployments that include SAS Visual Analytics, the service account for SAS LASR Analytic Server Monitor must be configured for passwordless SSH on each machine in the cluster. This is necessary to monitor hardware resources and processes for a distributed SAS LASR Analytic Server. This service account can be the same as the SAS installer account.

How to Set Up Passwordless SSH

You can use a point-and-click interface to generate SSH keys and configure them for passwordless SSH automatically for administrator accounts. See the [SAS High-Performance Computing Management Console: User's Guide](#).

Here are some tips:

- In the SAS High-Performance Computing Management Console, be sure to select the **Generate and Propagate SSH Keys** option on the Create User page. This ensures that passwordless SSH is configured correctly for the account.

- After you add user or group accounts to the machines in the cluster, you must restart the HDFS service if it is co-located. An error message such as the following indicates that a user is not recognized:

```
ERROR: host02.example.com (192.168.1.240) User does not belong to .
```

Generate SSH Keys Manually

The recommended method is to use the SAS High-Performance Computing Management Console to generate SSH keys (as described in the preceding topic).

If you must generate SSH keys manually (for example, for existing user IDs), use the following steps:

- 1 Generate a private/public key pair on a Linux system. Enter the following command to generate the keys without requiring a passphrase:

```
ssh-keygen -t rsa -P ""
```

- 2 After the keys are generated, if passwordless SSH is required, then add the public key to the list of authorized keys by entering this command on the command line:

```
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
```

- 3 Check permissions on the `.ssh` directory and the files in your `.ssh` directory. The directory must be readable and writable by you only. The `id_rsa` file must be readable by you only. To verify access, enter the following command, and check the results:

```
ls -asl ~/.ssh

4 drwx----- 2 datamgr datamgr 4096 Jan 23 10:27 . a
4 drwx----- 4 datamgr datamgr 4096 Jan 12 19:09 ..
4 -rw-r--r-- 1 datamgr datamgr 397 Jan 23 10:27 authorized_keys
4 -rw----- 1 datamgr datamgr 1675 Jan 23 10:00 id_rsa b
4 -rw-r--r-- 1 datamgr datamgr 397 Jan 13 10:00 id_rsa.pub
4 -rw-r--r-- 1 datamgr datamgr 1705 Jan 23 10:27 known_hosts
```

a The directory permissions for the `.ssh` directory indicate that access is denied for all users other than the directory owner.

b The `id_rsa` file is the private key. Read access and Write access are available to the file owner only.

Note: If the machines in the cluster are not configured to access the home directories for the users, create local home directories for the users. Copy the `.ssh` directory for each user to his or her local home directory. Make sure that the permissions are preserved.

About Passwordless SSH and Windows Clients

If you need to access a distributed SAS LASR Analytic Server from a Windows client, then you need to perform the following steps to copy your SSH keys to the Windows machine:

- 1 Determine your Windows home directory. Enter the following command in a command window:

```
echo %HOMEDRIVE%%HOMEPATH%
```

The results are typically something like `C:\Users\sasdemo`.

- 2 You can use Windows Explorer to drag-and-drop the `.ssh` directory from your UNIX home directory, or you can use a command like the following to copy it:

```
xcopy driverLetter:\.ssh\* "%HOMEDRIVE%%HOMEPATH%\.ssh" /s /i
```

These steps are typically necessary for deployments that use SAS Studio on a Windows client or SAS solutions that use Windows machines for the server tier.

Troubleshooting

If access problems occur, use the following steps to help diagnose any SSH configuration errors:

- 1 Impersonate the user or ask the user to perform the following command that requires passwordless SSH:

```
/opt/TKGrid/bin/simsh hostname
```

If each of the machines in the cluster responds with a host name, then no passwordless SSH configuration error exists.

- 2 As root, log on to one of the machines in the cluster and monitor the logon access:

```
tail -f /var/log/secure
```

- 3 Review the messages in the `/var/log/secure` file. The following example shows that the file system access permissions for `/home/sas` are not set correctly:

```
Mar 14 22:12:36 hostname sshd[11235]: pam_unix(sshd:session): session opened
for user root by (uid=0)
Mar 14 22:12:57 hostname sshd[11266]: Authentication refused: bad ownership or
modes for directory /home/sas
```


Recommended Reading

Here is the recommended reading list for this title:

- Configuration Guide for SAS Foundation for Microsoft Windows for x64.
- Configuration Guide for SAS Foundation for UNIX Environments.
- SAS/ACCESS for Relational Databases: Reference.
- SAS Deployment Wizard and SAS Deployment Manager: User's Guide.
- SAS Guide to Software Updates.
- SAS High-Performance Computing Management Console: User's Guide.
- SAS and Hadoop Technology: Deployment Scenarios
- SAS and Hadoop Technology: Overview.
- SAS In-Database Products: Administrator's Guide.
- SAS Intelligence Platform: Installation and Configuration Guide.
- SAS Intelligence Platform: Security Administration Guide.
- SAS LASR Analytic Server Reference Guide.
- SAS 9.4 Support for Hadoop.

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

SAS Books
SAS Campus Drive
Cary, NC 27513-2414
Phone: 1-800-727-0025
Fax: 1-919-677-4444
Email: sasbook@sas.com
Web address: sas.com/store/books

Glossary

browser

See [web browser](#).

co-located data provider

a distributed data source, such as SAS Visual Analytics Hadoop or a third-party vendor database, that has SAS High-Performance Analytics software installed on the same machines. The SAS software on each machine processes the data that is local to the machine or that the data source makes available as the result of a query.

data set

See [SAS data set](#).

data warehouse (warehouse)

a collection of pre-categorized data that is extracted from one or more sources for the purpose of query, reporting, and analysis. Data warehouses are generally used for storing large amounts of data that originates in other corporate applications or that is extracted from external data sources.

deployment plan

information about what software should be installed and configured on each machine in a SAS deployment. A deployment plan is stored in a plan.xml file.

encryption

the conversion of data by the use of algorithms or other means into an unintelligible form in order to secure data (for example, passwords) in transmission and in storage.

Extensible Markup Language (XML)

a markup language that structures information by tagging it for content, meaning, or use. Structured information contains both content (for example, words or numbers) and an indication of what role the content plays. For example, content in a section heading has a different meaning from content in a database table.

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 (HDFS)

a portable, scalable framework, written in Java, for managing large files as blocks of equal size. The files are replicated across multiple host machines in a Hadoop cluster in order to provide fault tolerance.

HDFS

See [Hadoop Distributed File System](#).

high-performance root node

See [root node](#).

identity

See [metadata identity](#).

Integrated Windows authentication (IWA)

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 (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.

IPv6

See [Internet Protocol Version 6](#).

IWA

See [Integrated Windows authentication](#).

JAR (Java Archive)

the name of a package file format that is typically used to aggregate many Java class files and associated metadata and resources (text, images, etc.) into one file to distribute application software or libraries on the Java platform.

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 Archive

See [JAR](#).

Java Database Connectivity (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.

Java Development Kit (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.

JDBC

See [Java Database Connectivity](#).

JDK

See [Java Development Kit](#).

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 (MPI)

a standardized and portable message-passing system that was designed to function on a wide variety of parallel computers. SAS Analytics applications implement MPI for use in high-performance computing environments.

metadata identity (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 (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.

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 (high-performance root node)

in a SAS High-Performance Analytics application, 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 (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 (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, 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 (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 server that provides access to SAS Foundation features such as the SAS programming language and SAS libraries.

SASHDAT file format

a SAS proprietary data format that is optimized for high performance and computing efficiency. For distributed servers, SASHDAT files are read in parallel. When used with the Hadoop Distributed File System (HDFS), the file takes advantage of data replication for fault-tolerant data access.

SASHOME directory

the location in a file system 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 that is installed on the same computer.

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 (SAS installation data file)

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

single sign-on (SSO)

an authentication model that enables users to access a variety of computing resources without being repeatedly prompted for their user IDs and passwords. For example, single sign-on can enable a user to access SAS servers that run on different platforms without interactively providing the user's ID and password for each platform. Single sign-on can also enable someone who is using one application to launch other applications based on the authentication that was performed when the user initially logged on.

SOE

See [software order email](#).

software order email (SOE)

an email message, sent to a customer site, that announces arrival of the software and describes the order. It explains the initial installation steps and might also contain instructions for using Electronic Software Delivery (ESD), if applicable.

spawner

See [object spawner](#).

SSO

See [single sign-on](#).

trusted user

a privileged service account that can act on behalf of other users on a connection to the metadata server.

unrestricted identity

a user or group that has all capabilities and permissions in the metadata environment due to membership in the META: Unrestricted Users Role (or listing in the adminUsers.txt file with a preceding asterisk).

update mode

an operating state of the SAS Deployment Wizard in which users are required to install software updates before they can perform any other deployment tasks. The SAS Deployment Wizard automatically goes into update mode when it determines that the current SAS order contains new versions or maintenance updates to the deployed products in a given SAS installation directory.

warehouse

See [data warehouse](#).

web application

an application that is accessed via a web browser over a network such as the Internet or an intranet. SAS web applications are Java Enterprise Edition (JEE) applications that are delivered via web application archive (WAR) files. The applications can depend on Java and non-Java web technologies.

web authentication

a configuration in which users of web applications and web services are verified at the web perimeter, and the metadata server trusts that verification.

web browser (browser)

a software application that is used to view web content, and also to download or upload information. The browser submits URL (Uniform Resource Locator) requests to a web server and then translates the HTML code into a visual display.

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](#).

XML

See [Extensible Markup Language](#).

Index

A

accounts
 See [user accounts](#)
 Authn::PAM PERL [20](#)
 authorized_keys file [26](#)

D

deployment
 overview [9](#)
 depot
 See [SAS Software Depot](#)

G

gridhosts file [20](#)
 groups
 setting up [13, 25, 53](#)

I

installation [1](#)

K

keys
 See [SSH public key](#)

M

middle tier shared key
 propagate [32](#)

O

operating system accounts
 See [user accounts](#)

P

perl-Net-SSLeay [20](#)

R

required user accounts [13, 25, 53](#)
 requirements, system [9](#)

S

SAS High-Performance Computing
 Management Console
 create user accounts [32](#)
 deployment [26](#)
 logging on [29](#)
 middle tier shared key [32](#)
 SAS High-Performance Computing
 Management Console server
 starting [27](#)
 SAS Software Depot [20](#)
 SAS system accounts [13, 25, 53](#)
 SAS Visual Analytics
 deploying [9](#)
 secure shell [20](#)
 JBoss Application Server public key
 [29](#)
 propagate keys [32](#)
 security
 passwordless SSH [93](#)
 server
 SAS High-Performance Computing
 Management Console [27](#)
 SSH
 See [secure shell](#)
 SSH public key
 JBoss Application Server [29](#)
 SSH public keys
 propagate [32](#)
 SSL [28](#)
 system requirements [9](#)

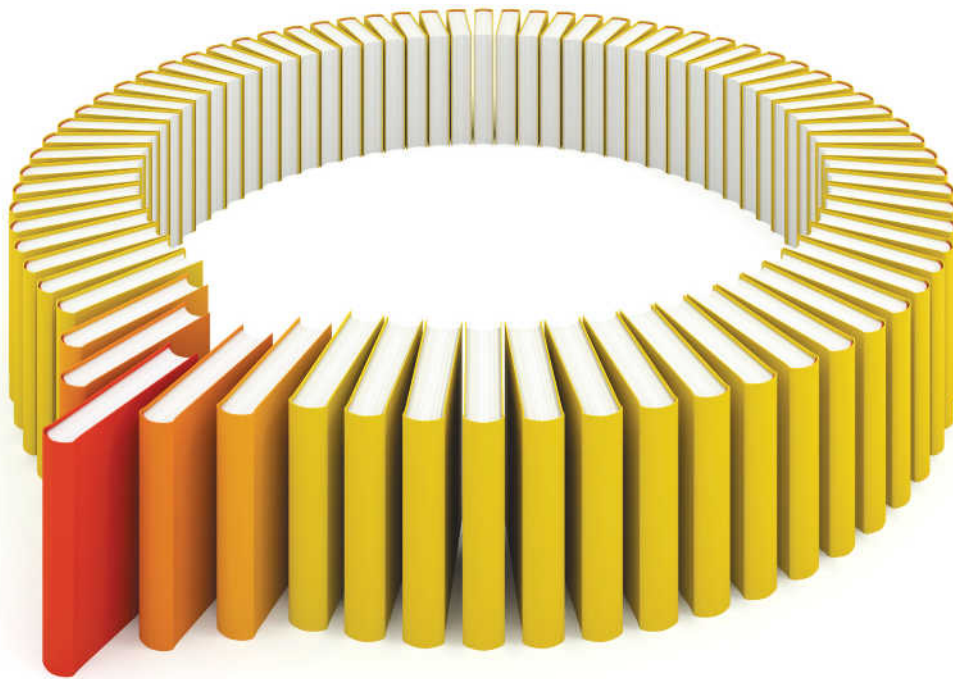
U

user accounts [13](#), [25](#), [53](#)

JBoss Application Server [29](#)

SAS system accounts [13](#), [25](#), [53](#)

setting up required accounts [13](#), [25](#),
[53](#)



Gain Greater Insight into Your SAS® Software with SAS Books.

Discover all that you need on your journey to knowledge and empowerment.

 support.sas.com/bookstore
for additional books and resources.


THE POWER TO KNOW.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. © 2013 SAS Institute Inc. All rights reserved. S107969US.0613

