# Contents

*Using This Book* ............................................................................................................. v

**Chapter 1 • Introduction to SAS and Hadoop Deployment Scenarios** ................................................. 1
  Deploying SAS with Hadoop ......................................................................................... 1
  How to Use This Guide ................................................................................................. 3
  Tips Before You Deploy ............................................................................................... 3
  Supported Hadoop Distributions .................................................................................. 5

**Chapter 2 • Scenario for SAS Data Loader for Hadoop** ......................................................... 7
  Overview ..................................................................................................................... 7
  Step 1: Install and Configure Software on the Hadoop Cluster ..................................... 9
  Step 2: Deploy the vApp and Configure Directives ..................................................... 10

**Chapter 3 • Scenarios for In-Memory Analytics** ........................................................................ 11
  Overview ..................................................................................................................... 12
  Scenario 1: Deploy In-Memory Analytics with SAS
    High-Performance Deployment of Hadoop .................................................................. 19
  Scenario 2: Deploy In-Memory Analytics on Your Hadoop Cluster ............................. 23
  Scenario 3: Configure Remote Access to Hadoop ..................................................... 28

*Recommended Reading* .................................................................................................. 31

*Index* .................................................................................................................................. 33
Audience

This document is for anyone who is interested in learning how to deploy SAS software with Hadoop. Examples include a Hadoop administrator or IT administrator who will install SAS software to work with Hadoop, or a SAS representative who helps customers deploy software.
Using This Book
Introduction to SAS and Hadoop Deployment Scenarios

Deploying SAS with Hadoop

SAS has long supported many different data sources, and Hadoop is no exception. Currently, more than twenty SAS products, solutions, and technology packages interact with Hadoop. Each SAS technology provides different functionality—from accessing and managing Hadoop data to executing in-memory analytics with data in Hadoop.

Because SAS provides multiple options for accessing and processing data in Hadoop, consider the following approaches to deployment:

- You can configure SAS software that enables processing in your Hadoop cluster. To do this, you deploy the SAS in-database deployment package for Hadoop (SAS Embedded Process) on the same nodes of your Hadoop cluster. Together with
SAS/ACCESS Interface to Hadoop, SAS Embedded Process enables analysts to run programs in Hadoop. This approach eliminates data movement, as SAS performs the analytics where the data is stored and uses the distributed, parallel processing architecture of Hadoop for improved performance.

SAS products that can take advantage of data management in the Hadoop cluster include SAS Data Loader for Hadoop, SAS Scoring Accelerator for Hadoop, and SAS Code Accelerator for Hadoop.

This guide provides a deployment scenario for the SAS Data Loader for Hadoop. For more information, see Chapter 2, “Scenario for SAS Data Loader for Hadoop,” on page 7.

You can deploy an in-memory analytics environment to work with Hadoop. This approach provides the greatest potential for the fastest analytics on very large data sets from Hadoop.

The SAS software that makes up an in-memory analytics environment is SAS High-Performance Analytics and the SAS LASR Analytic Server. Products that can take advantage of this environment include SAS Visual Analytics and SAS Visual Statistics. This guide provides different deployment scenarios for in-memory analytics. For more information, see Chapter 3, “Scenarios for In-Memory Analytics,” on page 11.

You can configure SAS for basic access to Hadoop so that users can access data in Hadoop, just as they would with any other data source. This approach supports the most products provided by SAS, and any tools that you already have in place can access and manage data from Hadoop.

Because configuring basic access to Hadoop includes post-installation configuration steps for both Base SAS and SAS/ACCESS Interface to Hadoop, this document provides no scenario and no additional planning information. Instead, you can find instructions in SAS Hadoop Configuration Guide for Base SAS and SAS/ACCESS.

**TIP** For more information about how SAS and Hadoop work together, see *SAS and Hadoop Technology: Overview*. 
How to Use This Guide

Use each scenario as a roadmap for deployment. Each scenario includes a summary of what is deployed, an overview of steps to complete the deployment, and links to SAS product documentation where you can find detailed instructions.

Each SAS solution, product, or technology has installation documentation and information to help you configure SAS software with Hadoop. Because some SAS deployments with Hadoop require multiple software components, information for how to best deploy the software is required. To be successful, make sure that you have access to the product documentation for the software that you want to deploy.

Tips Before You Deploy

Here are a few considerations before you get started.

Understand Your Hadoop Environment

To deploy SAS software with Hadoop successfully, consider the following tips:

- Gain working knowledge of the Hadoop distribution that you are using (for example, Cloudera CDH or Hortonworks Data Platform). Make sure you have working knowledge of the Hadoop Distributed File System (HDFS), MapReduce 1, MapReduce 2, YARN, Hive, and HiveServer2 services. Review your YARN configuration. For more information, see the Apache website or the vendor’s website.

- Ensure that the HCatalog, HDFS, Hive, MapReduce, Oozie, Sqoop, and YARN services are running on the Hadoop cluster. SAS software uses these various services and you need to ensure that the appropriate JAR files are gathered during the configuration.

- Know the location of the MapReduce home.
Know the host name of the Hive server and the name of the NameNode.

Determine where the HDFS and Hive servers are running. If the Hive server is not running on the same machine as the NameNode, note the server and port number of the Hive server for future configuration.

Request permission to restart the MapReduce service.

Understand and verify your Hadoop user authentication.

Understand Kerberos or another security protocol for data security. Verify that you can connect to your Hadoop cluster (HDFS and Hive) from your client machine outside the SAS environment with your defined security protocol.

Note: The scenarios in this document assume that Kerberos has been enabled and both SAS software and the Hadoop cluster are configured as part of the same Kerberos realm.

Understand the SAS Intelligence Platform

The SAS Intelligence Platform provides the architecture for data management, business intelligence, and analytics. When you deploy SAS software for Hadoop, it is important to understand the different computing tiers and servers that the architecture comprises.

For more information, see SAS *Intelligence Platform: Overview*.

A SAS Software Depot Is Required

Before you can install SAS software, you must have a SAS Software Depot. The SAS Software Depot consists of a collection of SAS installation files that represent one or more software orders. The depot contains the SAS Deployment Wizard executable, one or more deployment plans (if you have a planned deployment), a SAS installation data file, and order data and product data.
Supported Hadoop Distributions

SAS supports commercial Hadoop distributions from Cloudera, Hortonworks, IBM BigInsights, MapR, and Pivotal.

For more information about the supported distributions, see SAS 9.4 Support for Hadoop. In addition, see the full product documentation or system requirements documentation for SAS products and technologies.
Chapter 1 / Introduction to SAS and Hadoop Deployment Scenarios
Scenario for SAS Data Loader for Hadoop

Overview

You can deploy SAS Data Loader for Hadoop to take advantage of the data management capabilities of Hadoop. This chapter provides an overview of SAS Data Loader for Hadoop and a roadmap for deploying software.

About SAS Data Loader for Hadoop

SAS Data Loader for Hadoop is a software offering that makes it easier to move, cleanse, and analyze data in Hadoop. SAS Data Loader for Hadoop provides a set of directives or wizards that help business users and data scientists perform the following tasks:

- copy data to and from Hadoop using parallel bulk data transfer
perform data integration, data quality, and data preparation tasks within Hadoop, without writing complex MapReduce code or asking for outside help

- minimize data movement for increased scalability, governance, and performance
- load data in memory to prepare it for high-performance reporting, visualization, or analytics

What Gets Deployed

To use SAS Data Loader for Hadoop, you deploy the software components shown in the following table.

Table 2.1  Software Included with SAS Data Loader for Hadoop

<table>
<thead>
<tr>
<th>Software Branch Name</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAS/ACCESS Interface to Hadoop</td>
<td>The access engine for interfacing with Hadoop. SAS/ACCESS Interface to Hadoop is a prerequisite for installing SAS Embedded Process for Hadoop.</td>
</tr>
</tbody>
</table>
| SAS Embedded Process for Hadoop (in-database deployment package for Hadoop) | The software that enables the following technologies for SAS Data Loader for Hadoop:  
  - SAS Data Quality Accelerator for Hadoop, the component that is required to run the Cleanse Data in Hadoop directive in the Hadoop cluster.  
  - SAS Quality Knowledge Base (QKB), which is a collection of files that store data and logic that support data management operations.  
  - SAS In-Database Code Accelerator for Hadoop, a component that supports running DS2 threaded programs in Hadoop. |
| vApp for SAS Data Loader for Hadoop (client) | The virtual machine that business users run to interface with SAS Data Loader for Hadoop. The vApp is a complete and isolated operating environment that business users configure in a supported hypervisor. |
Step 1: Install and Configure Software on the Hadoop Cluster

The following steps should be performed by a Hadoop or systems administrator.

**Note:** The steps for this scenario are for deploying SAS Data Loader 2.3 for Hadoop.

**Table 2.2  Software Deployed on the Hadoop Cluster**

<table>
<thead>
<tr>
<th>Overview of Steps</th>
<th>Documentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Install SAS/ACCESS Interface to Hadoop.</td>
<td><em>SAS Hadoop Configuration Guide for Base SAS and SAS/ACCESS</em></td>
</tr>
<tr>
<td>2. Deploy SAS Embedded Process</td>
<td><em>SAS In-Database Products: Administrator's Guide</em></td>
</tr>
<tr>
<td></td>
<td>See “Part 3 — Administrator’s Guide for SAS Data Loader for Hadoop.”</td>
</tr>
</tbody>
</table>

**Note:**

- If business users require the Cleanse Data in Hadoop directive, then follow instructions in “Chapter 8 — SAS Data Quality Accelerator for Hadoop.”

- If your enterprise uses Kerberos security, additional configuration is required on the Hadoop cluster. For more information, see “Chapter 9 — Security.”
### Overview of Steps

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.</td>
<td>Provide information to each person who will deploy the vApp for SAS Data Loader for Hadoop. <strong>Note:</strong> This information includes the host name and port number of the Hadoop server, the location of Hadoop configuration files, and Kerberos settings to run SAS Data Loader for Hadoop in secure mode. Each person deploying a vApp must have the correct information.</td>
</tr>
</tbody>
</table>

### Step 2: Deploy the vApp and Configure Directives

To deploy SAS Data Loader for Hadoop on each client host, a user must set up and run the vApp in a supported hypervisor. Part of the configuration is to enter information that is provided by the Hadoop administrator.

### Table 2.3 vApp Setup and Configuration

<table>
<thead>
<tr>
<th>Overview of Steps</th>
<th>Documentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Set up and run the vApp.</td>
<td><strong>SAS Data Loader for Hadoop: vApp Deployment Guide</strong></td>
</tr>
<tr>
<td>2. Configure directives and global settings. Additional configuration might be required for global settings or for working with the different directives.</td>
<td>See the “Get Started” section in <strong>SAS Data Loader for Hadoop: User’s Guide</strong>.</td>
</tr>
</tbody>
</table>
Scenarios for In-Memory Analytics

Overview  ................................................................. 12
  About In-Memory Analytics .................................. 12
  Deployment Considerations ................................. 13
  Core Software ..................................................... 15
  Supporting Software ........................................... 16
  Hadoop Distributions .......................................... 17
  SAS Products That Take Advantage of In-Memory Analytics .... 18

Scenario 1: Deploy In-Memory Analytics with
SAS High-Performance Deployment of Hadoop .............. 19
  What Gets Deployed ............................................ 19
  Step 1: Prepare Your Machines .............................. 20
  Step 2: Deploy SAS High-Performance
    Computing Management Console and Create
    User Accounts ................................................. 21
  Step 3: Deploy SAS High-Performance
    Deployment of Hadoop ..................................... 21
  Step 4: Install and Configure the SAS High-
    Performance Analytics Environment ................... 22

Scenario 2: Deploy In-Memory Analytics on Your
Hadoop Cluster ..................................................... 23
  What Gets Deployed ............................................ 23
  Step 1: Prepare Machines on the Hadoop Cluster ....... 24
Overview

You can deploy software that enables SAS software to process and analyze data in a distributed, in-memory environment. Because different environments are possible, this chapter provides information to help you understand the SAS software that enables in-memory analytics, and each scenario provides a roadmap for deploying the software.

About In-Memory Analytics

A key to understanding the advantages of in-memory analytics is to understand how data is staged and where analytics takes place. To process data, SAS loads data from Hadoop to the in-memory environment. The in-memory environment then performs the analysis and only the results are returned to the SAS server or SAS client that submitted the request.

The in-memory SAS software consists of the SAS High-Performance Analytics environment and SAS LASR Analytic Server. The set of connected machines where SAS in-memory software is deployed and where SAS processes the data is referred to as the analytics cluster. When it is deployed, the analytics cluster reduces processing...
time and brings SAS High-Performance Analytics to data volumes that exceed the memory capacity of a single machine.

To read more about the advantages of using in-memory analytics, see SAS and Hadoop Technology: Overview.

**Deployment Considerations**

Different deployments are possible for an in-memory environment.

- In-memory analytics software can be deployed with SAS High-Performance Deployment of Hadoop. An advantage to this scenario is that the Hadoop distribution provided by SAS can be used to stage data in HDFS as SASHDAT tables. The SASHDAT file format is memory efficient, supports SAS formats, and is optimal for use with in-memory processing.

  For more information, see “Scenario 1: Deploy In-Memory Analytics with SAS High-Performance Deployment of Hadoop” on page 19.

- You can co-locate the in-memory environment on a commercial Hadoop distribution. You might want to do this if you have experience with a commercial distribution or prefer the management interfaces that it provides. After you configure the commercial distribution, all the benefits of the SASHDAT file format that are ascribed to SAS High-Performance Deployment of Hadoop are available. If you already have a Hadoop cluster and the resource demands leave capacity for SAS software, you can consider deploying SAS software on the same machines.

  For more information, see “Scenario 2: Deploy In-Memory Analytics on Your Hadoop Cluster” on page 23.

- You can configure your analytics cluster to access a remote Hadoop cluster. In this scenario, minimal SAS software is installed in the Hadoop cluster. Specifically, SAS Embedded Process is deployed on the nodes of the Hadoop cluster to handle asymmetric, distributed workloads from remote Hadoop to the SAS analytics cluster. In this scenario, it is recommended that you deploy the in-memory analytics software with SAS High-Performance Deployment of Hadoop to stage data in HDFS as SASHDAT tables.
For more information, see “Scenario 3: Configure Remote Access to Hadoop” on page 28.

**Note:** A co-located deployment consists of SAS in-memory analytics software that is installed on the same nodes as a distribution of Hadoop. As described previously, the Hadoop distribution can be the SAS High-Performance Deployment of Hadoop or a commercial Hadoop distribution. Here are two ways a co-located deployment can help:

- To use SASHDAT tables, a co-located deployment is required. An important feature of a co-located deployment is the ability to stage data on the analytics cluster. The preferred format is a SASHDAT table, because it takes advantage of the redundancy and highly available features of HDFS. SAS LASR Analytic Server and high-performance procedures can read and write SASHDAT tables in parallel at impressive speeds.

- If you use SAS LASR Analytic Server, a co-located deployment is highly recommended. A co-located deployment enables you to read data from operational systems and stage it so that SAS can analyze it on the same machines. One of the most powerful benefits of SAS LASR Analytic Server is the ability to read data in parallel from a co-located data provider. A co-located deployment provides optimal performance when using the in-memory processing.
Core Software

The primary SAS software that enables distributed, in-memory computing is SAS High-Performance Analytics and SAS LASR Analytic Server.

Table 3.1  Software for In-Memory Analytics

<table>
<thead>
<tr>
<th>Software</th>
<th>Details</th>
</tr>
</thead>
</table>
| SAS High-Performance Analytics environment (also known as SAS High-Performance Analytics node installation) | - Performs analytic tasks in a high-performance environment that is characterized by massively parallel processing (MPP). After you deploy it, a root-and-worker architecture is established for running distributed, high-performance analytics.  
  - TKGrid is the primary software installed on each node to provide the SAS High-Performance Analytics environment. |
| SAS LASR Analytic Server                                   | - A scalable, analytic platform that provides a secure, multi-user environment for concurrent access to in-memory data.  
  - Provides the ability to load Hadoop data into memory and perform a variety of distributed processing, exploratory analyses, analytic calculations, and more. |
Supporting Software

SAS provides additional software that can enable access, enhance performance, and facilitate administration of the in-memory environment.

Table 3.2  Supporting Software for In-Memory Analytics

<table>
<thead>
<tr>
<th>Software</th>
<th>Details</th>
</tr>
</thead>
</table>
| SAS Embedded Process (in-database deployment package for Hadoop) | - Recommended software that enables reading and writing data to HDFS in parallel for SAS High-Performance Analytics.  
                   | - For environments that access Hadoop remotely, SAS Embedded Process is required to handle asymmetric, parallel loads between HDFS and the analytics cluster.  
                   | - SAS Embedded Process is not supported for environments that use SAS High-Performance Deployment of Hadoop.                           |
| SAS/ACCESS Interface to Hadoop                                | - The required access engine that enables SAS software to interface with Hadoop.  
                   | - For environments that access Hadoop remotely, SAS/ACCESS Interface to Hadoop works with SAS Embedded Process to read data from the Hadoop cluster. |
| SAS High-Performance Computing Management Console             | - A console that provides an easy-to-use interface for performing administrative tasks in the analytics cluster.  
                   | - This software is optional.                                                                                                        |

**Note:** A SAS client is required to submit programs to the analytics cluster, and SAS/ACCESS Interface to Hadoop must be installed on the same machine as the SAS client. Data scientists, analytic experts, and other users interface with the SAS client.
software to write SAS programs and submit them to the SAS High-Performance Analytics environment. An example of a SAS client is SAS Studio.

## Hadoop Distributions

The Hadoop distributions mentioned in this section are shown in the following table.

### Table 3.3  Supporting Software for In-Memory Analytics

<table>
<thead>
<tr>
<th>Software</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAS High-Performance Deployment of Hadoop</td>
<td>A Hadoop distribution provided by SAS as a convenience for deploying co-located Hadoop with your analytics cluster. This software is optional and is not intended to replace a commercial distribution of Hadoop.</td>
</tr>
<tr>
<td></td>
<td>Includes the Apache Hadoop framework, which includes Hadoop Common, HDFS, Hadoop YARN, and Hadoop MapReduce. Also includes JAR files from SAS that provide support for SASHDAT tables.</td>
</tr>
<tr>
<td>Commercial Hadoop distributions</td>
<td>The collection of Hadoop components (such as HDFS, Hive, and MapReduce) that is provided by a vendor.</td>
</tr>
<tr>
<td></td>
<td>For more information about the supported distributions, see <a href="#">SAS 9.4 Support for Hadoop</a>. In addition, see the full product documentation or system requirements documentation for SAS products and technologies.</td>
</tr>
</tbody>
</table>
SAS Products That Take Advantage of In-Memory Analytics

If you intend to use one or more of the following products, consider deploying an in-memory environment.

Table 3.4  SAS Products That Take Advantage of In-Memory Analytics

<table>
<thead>
<tr>
<th>Software</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAS Visual Analytics</td>
<td>Use SAS Visual Analytics to explore large volumes of data very quickly to identify patterns and trends and to identify opportunities for further analysis. Visual Analytics provides an easy-to-use, web-based interface for running analytics.</td>
</tr>
<tr>
<td>SAS In-Memory Statistics</td>
<td>Use SAS In-Memory Statistics to perform analytical data preparation, variable transformations, exploratory analysis, statistical modeling and machine-learning techniques, integrated modeling comparison, and model scoring. Data scientists can access data from a variety of sources and use an interactive programming interface to access data in memory. The IMSTAT procedure enables in-memory analytics on the data. SAS LASR Analytic Server holds the data in memory and performs complex analytics.</td>
</tr>
<tr>
<td>SAS High-Performance Risk</td>
<td>These products, which provide multiple features for modeling, are collectively known as the SAS High-Performance Analytics products.</td>
</tr>
<tr>
<td>SAS High-Performance Data Mining</td>
<td></td>
</tr>
<tr>
<td>SAS High-Performance Econometrics</td>
<td></td>
</tr>
<tr>
<td>SAS High-Performance Optimization</td>
<td></td>
</tr>
<tr>
<td>SAS High-Performance Statistics</td>
<td></td>
</tr>
<tr>
<td>SAS High-Performance Text-Mining</td>
<td></td>
</tr>
</tbody>
</table>
Scenario 1: Deploy In-Memory Analytics with SAS High-Performance Deployment of Hadoop

What Gets Deployed

In this scenario, the SAS in-memory analytics environment is installed on the same nodes with SAS High-Performance Deployment of Hadoop.

The following table shows nodes and roles, as well as the locations where analytics software is installed.

Table 3.5  Roles and Software for Co-Located SAS High-Performance Deployment of Hadoop

<table>
<thead>
<tr>
<th>On the Root Node</th>
<th>On Each Worker Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAS Software</td>
<td>SAS Software</td>
</tr>
<tr>
<td>- NameNode: SAS High-Performance Deployment of Hadoop</td>
<td>- DataNode: SAS High-Performance Deployment of Hadoop</td>
</tr>
<tr>
<td>- Root Node: SAS High-Performance Analytics Environment</td>
<td>- Worker Node: SAS High-Performance Analytics Environment</td>
</tr>
<tr>
<td>- SAS High-Performance Management Console</td>
<td></td>
</tr>
</tbody>
</table>

Note: Both the Hadoop NameNode and the root node for SAS High-Performance Analytics are on the same machine. The root node takes on the role of distributing and coordinating the workload to the worker nodes.
## Step 1: Prepare Your Machines

An important step for a successful deployment is to ensure that the machines are configured appropriately before you deploy SAS High-Performance Deployment of Hadoop and the SAS High-Performance Analytics environment.

### Table 3.6  Configuration to Prepare Your Machines

<table>
<thead>
<tr>
<th>Overview of Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Note:</strong> To find the detailed information for the following steps, see “Chapter 2 — Preparing Your System to Deploy the SAS High-Performance Analytics Infrastructure” in <em>SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide</em>.</td>
</tr>
<tr>
<td>1.</td>
</tr>
<tr>
<td>2.</td>
</tr>
<tr>
<td>3.</td>
</tr>
</tbody>
</table>
| 4. | Prepare to install SAS High-Performance Computing Management Console.  
**Note:** SAS High-Performance Computing Management Console is an optional installation. In this scenario, the console is installed to highlight the features that it provides. |
| 5. | Prepare to deploy Hadoop. |
| 6. | Prepare to deploy the SAS High-Performance Analytics environment. |
| 7. | Read information about recommended data names. |
| 8. | Make sure the SAS Software Depot has been created and is available to the root node. |
| 9. | Make sure the installer is root. Also, review the user account and directory recommendations. |
| 10. | Make sure Java software is installed. Each machine must have a Java Runtime Environment (JRE) or Java Development Kit (JDK) installed. |
Overview of Steps

11. Understand user account requirements and umask settings for deploying and running the SAS High-Performance Analytics environment.

12. Recommended: Record each SAS port in `/etc/services`.

Step 2: Deploy SAS High-Performance Computing Management Console and Create User Accounts

When users interact with the analytics cluster, user accounts with passwordless SSH are required to start and stop SAS LASR Analytic Server and to run programs in the analytics cluster. The SAS High-Performance Computing Management Console is designed to ease the creation of user accounts that require passwordless SSH.

Table 3.7  Installing and Configuring the Console and Creating Users

<table>
<thead>
<tr>
<th>Overview of Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1.</strong> Install the SAS High-Performance Management Console by using RPM or tar.</td>
</tr>
<tr>
<td><strong>2.</strong> Configure the SAS High-Performance Management Console.</td>
</tr>
<tr>
<td><strong>3.</strong> Create user accounts and propagate the SSH key for each account.</td>
</tr>
</tbody>
</table>

Step 3: Deploy SAS High-Performance Deployment of Hadoop

SAS High-Performance Deployment of Hadoop is deployed on all machines where you plan to deploy and run the SAS High-Performance Analytics environment.
Table 3.8  Installing SAS High-Performance Deployment of Hadoop

Overview of Steps

Note: To find the detailed information for the following steps, see “Chapter 4 — Deploying Co-Located Hadoop” in SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide. Look for step-by-step instructions in the section “Deploying SAS High-Performance Deployment of Hadoop.”

1. Install SAS High-Performance Deployment of Hadoop.

2. Perform manual steps after installation, including configuration changes for Kerberos.

3. Validate the SAS High-Performance Deployment of Hadoop.

Step 4: Install and Configure the SAS High-Performance Analytics Environment

Deploying the SAS High-Performance Analytics environment requires installing and configuring components on the machine that will act as the root node, and then on the remaining worker nodes. Refer to the following table for an overview of steps.

Table 3.9  Deploying the SAS High-Performance Analytics Environment

Overview of Steps

Note: To find the detailed information for the following steps, see “Chapter 5 — Deploying the SAS High-Performance Analytics Environment” in SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide.

1. Install the analytics environment by running the TKGrid and TKTGDat shell scripts.

2. Validate the environment by invoking the simsh or MPI command.

3. (Optional) Customize resource settings for your site. To manage the performance, you can set limits on processes that run in the analytics cluster, and you can control how much memory is requested by programmers.
Note: You can configure the analytics environment for SASHDAT encryption. Because this scenario does not cover this feature, see “Chapter 5 — Deploying the SAS High-Performance Analytics Environment” in SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide for more information.

Scenario 2: Deploy In-Memory Analytics on Your Hadoop Cluster

What Gets Deployed

In this scenario, the software that makes up the in-memory environment is deployed on the same nodes with your commercial Hadoop distribution.

The following table shows nodes and roles, as well as locations where analytics software is installed.

Table 3.10  Roles and Software for a Co-Located Commercial Hadoop Distribution

<table>
<thead>
<tr>
<th>On the Root Node</th>
<th>On Each Worker Node</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Commercial Hadoop</td>
<td>Commercial Hadoop</td>
</tr>
<tr>
<td>NameNode</td>
<td>DataNode</td>
</tr>
<tr>
<td>SAS Software</td>
<td>SAS Software</td>
</tr>
<tr>
<td>SAS Root Node: SAS High-Performance Analytics Environment</td>
<td>Worker Node: SAS High-Performance Node</td>
</tr>
<tr>
<td>SAS High-Performance Management Console (optional)</td>
<td>SAS Embedded Process</td>
</tr>
<tr>
<td>SAS Embedded Process</td>
<td></td>
</tr>
</tbody>
</table>

Note: Both the Hadoop NameNode and the root node for SAS High-Performance Analytics are installed on the same machine. The root node takes on the role of distributing and coordinating the workload to the worker nodes.
Step 1: Prepare Machines on the Hadoop Cluster

An important step for a successful deployment is to ensure that the machines are configured appropriately before you deploy SAS High-Performance Deployment of Hadoop and the SAS High-Performance Analytics environment.

Table 3.11 Configuration to Prepare Your Machines

<table>
<thead>
<tr>
<th>Overview of Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Note:</strong> To find the detailed information for the following steps, see “Chapter 2 — Preparing Your System to Deploy the SAS High-Performance Analytics Infrastructure” in SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide.</td>
</tr>
<tr>
<td>1. Configure system settings on all nodes.</td>
</tr>
<tr>
<td>2. Create the <code>/etc/gridhosts</code> file, and list all nodes that will run the analytics cluster. During a later step in this scenario, you will reference this file so that the analytics cluster is installed on the relevant nodes.</td>
</tr>
<tr>
<td>3. Prepare each machine for Kerberos.</td>
</tr>
</tbody>
</table>
| 4. Prepare to install SAS High-Performance Computing Management Console.  
  **Note:** SAS High-Performance Computing Management Console is an optional installation. In this scenario, the console is installed to highlight the features that it provides. |
| 5. Prepare to deploy Hadoop. |
| 6. Prepare to deploy the SAS High-Performance Analytics environment. |
| 7. Read information about recommended data names. |
| 8. Make sure the SAS Software Depot has been created and is available to the root node. |
| 9. Make sure the installer is root. Also, review the user account and directory recommendations. |
Overview of Steps

10. Make sure Java software is installed. Each machine must have a Java Runtime Environment (JRE) or Java Development Kit (JDK) installed.

11. Understand user account requirements and umask settings for deploying and running the SAS High-Performance Analytics environment.

12. Recommended: Record each SAS port in /etc/services.

Step 2: Deploy SAS High-Performance Computing Management Console and Create User Accounts

When users interact with the analytics cluster, user accounts with passwordless SSH are required to start and stop SAS LASR Analytic Server and to run programs in the analytics cluster. The SAS High-Performance Computing Management Console is designed to ease the creation of user accounts that require passwordless SSH.

Table 3.12 Installing and Configuring the Console and Creating Users

Overview of Steps

Note: To find the detailed information for the following steps, see “Chapter 3 — Deploying SAS High-Performance Computing Management Console” in SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide.

1. Install the SAS High-Performance Management Console by using RPM or tar.

2. Configure the SAS High-Performance Management Console.

3. Create user accounts and propagate the SSH key for each account.
Step 3: Configure the Hadoop Cluster

For most commercial Hadoop distributions, you perform configuration steps on every machine in the Hadoop cluster. These steps can include setting environment variables, propagating the sas.lasr.jar and sas.lar.hadoop.jar files, and additional configuration.

Table 3.13  Configure the Hadoop Cluster

<table>
<thead>
<tr>
<th>Overview of Steps</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Note:</strong> To find the detailed information for the following steps, see “Chapter 4 — Deploying Co-Located Hadoop” in SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide. Look for step-by-step instructions for your specific Hadoop distribution in the section “Configuring Existing Hadoop Clusters.”</td>
<td></td>
</tr>
</tbody>
</table>

1. Read about the prerequisites for existing Hadoop clusters.

2. Follow steps specific to your existing implementation of Hadoop:
   - Cloudera CDH
   - Hortonworks HDP
   - IBM BigInsights
   - MapR Distribution
   - Pivotal HD

3. Create user accounts and propagate the SSH key for each account.

Step 4: Install and Configure the SAS High-Performance Analytics Environment

Deploying the SAS High-Performance Analytics environment requires installing and configuring components on the machine that will act as the root node, and then on the remaining worker nodes. Refer to the following table for an overview of steps.
Overview of Steps

1. Install the analytics environment by running the TKGrid, TKTGDat, and TKGrid_REP shell scripts.
   \textbf{Note:} TKGrid\_REP is required to work with the SAS Embedded Process for parallel reading and writing of non-SASHDAT Hadoop data.

2. Validate the environment by invoking the \texttt{simsh} or MPI command.

3. (Optional) Customize resource settings for your site. To manage the performance of the SAS High-Performance Analytics environment, you can set limits on processes that run in the analytics cluster, and you can control how much memory is requested by programmers.

\textbf{Note:} You can configure the analytics environment for SASHDAT encryption. Because this scenario does not cover this feature, see “Chapter 5 — Deploying the SAS High-Performance Analytics Environment” in \textit{SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide} for more information.

**Step 5: Deploy SAS Embedded Process**

SAS Embedded Process is required for in-database functionality with the Hadoop cluster, including scoring acceleration, code acceleration, and the SQL pass-through facility. SAS Embedded Process provides parallel data loading to SAS LASR Analytic Server functionality for Hive and Impala tables and for SPD Engine and SPD Server tables.

SAS Embedded Process is part of the in-database deployment package for Hadoop. For more information, see “Part 2 — Administrator’s Guide for Hadoop (In-Database Deployment Package)” in \textit{SAS In-Database Products: Administrator’s Guide}. 

---

**Table 3.14 Deploying the SAS High-Performance Analytics Environment**

<table>
<thead>
<tr>
<th>Overview of Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textbf{Note:} To find the detailed information for the following steps, “Chapter 5 — Deploying the SAS High-Performance Analytics Environment” in \textit{SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide}.</td>
</tr>
</tbody>
</table>

1. Install the analytics environment by running the TKGrid, TKTGDat, and TKGrid\_REP shell scripts. 
   \textbf{Note:} TKGrid\_REP is required to work with the SAS Embedded Process for parallel reading and writing of non-SASHDAT Hadoop data.

2. Validate the environment by invoking the \texttt{simsh} or MPI command.

3. (Optional) Customize resource settings for your site. To manage the performance of the SAS High-Performance Analytics environment, you can set limits on processes that run in the analytics cluster, and you can control how much memory is requested by programmers.

\textbf{Note:} You can configure the analytics environment for SASHDAT encryption. Because this scenario does not cover this feature, see “Chapter 5 — Deploying the SAS High-Performance Analytics Environment” in \textit{SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide} for more information.

**Step 5: Deploy SAS Embedded Process**

SAS Embedded Process is required for in-database functionality with the Hadoop cluster, including scoring acceleration, code acceleration, and the SQL pass-through facility. SAS Embedded Process provides parallel data loading to SAS LASR Analytic Server functionality for Hive and Impala tables and for SPD Engine and SPD Server tables.

SAS Embedded Process is part of the in-database deployment package for Hadoop. For more information, see “Part 2 — Administrator’s Guide for Hadoop (In-Database Deployment Package)” in \textit{SAS In-Database Products: Administrator’s Guide}. 

---

**Scenario 2: Deploy In-Memory Analytics on Your Hadoop Cluster**

27
Note: A prerequisite to installing SAS Embedded Process is to configure SAS/ACCESS Interface to Hadoop. For more information, see *SAS Hadoop Configuration Guide for Base SAS and SAS/ACCESS*.

---

**Scenario 3: Configure Remote Access to Hadoop**

**What Gets Deployed**

In this scenario, you deploy software on your existing commercial Hadoop cluster. This deployment enables a remote, parallel connection between a Hadoop cluster and a set of machines that is dedicated to running in-memory analytics. For this scenario, SAS/ACCESS Interface to Hadoop is installed on a SAS client machine that is remote from the Hadoop cluster and the in-memory analytics cluster.

The following table shows SAS Embedded Process software installed in the Hadoop cluster.

*Table 3.15  SAS Embedded Software Installed on the Hadoop Cluster*

<table>
<thead>
<tr>
<th>On the Root Node</th>
<th>On Each Worker Node</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Diagram" /></td>
<td><img src="image" alt="Diagram" /></td>
</tr>
<tr>
<td>Commercial Hadoop</td>
<td>Commercial Hadoop</td>
</tr>
<tr>
<td>■ NameNode</td>
<td>■ DataNode</td>
</tr>
<tr>
<td>SAS Software</td>
<td>SAS Software</td>
</tr>
<tr>
<td>■ SAS Embedded Process</td>
<td>■ SAS Embedded Process</td>
</tr>
</tbody>
</table>
Step 1: Install SAS/ACCESS Interface to Hadoop

In this scenario, SAS/ACCESS Interface to Hadoop is installed on a SAS client that is remote from the root node of the analytics cluster. For more information about configuring SAS/ACCESS Interface to Hadoop, see SAS Hadoop Configuration Guide for Base SAS and SAS/ACCESS.

Step 2: Deploy SAS Embedded Process in Your Hadoop Cluster

Make sure that SAS Embedded Process is deployed on each node in your Hadoop cluster.

For more information, see “Part 2 — Administrator’s Guide for Hadoop (In-Database Deployment Package)” in SAS In-Database Products: Administrator’s Guide.

Step 3: Configure the Analytics Environment for a Remote Parallel Connection

Make sure that TKGrid_REP is configured across all nodes in the in-memory analytics cluster. TKGrid_REP is a configuration of TKGrid that enables support for remote access to Hadoop.

Table 3.16  Deploying the SAS High-Performance Analytics Environment

<table>
<thead>
<tr>
<th>Overview of Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Note:</strong> To find the detailed information for the following steps, see “Chapter 6 — Configuring the Analytics Environment for a Remote Parallel Connection” in SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide.</td>
</tr>
</tbody>
</table>

1. Prepare for a remote parallel connection.

2. Read how the configuration script works.

3. Run the TKGrid_REP script to configure access.
Recommended Reading

Here is the recommended reading list for this title:

- SAS and Hadoop Technology: Overview
- SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide
- SAS Visual Analytics: Installation and Configuration Guide (Distributed SAS LASR)
- SAS In-Database Products: Administrator's Guide
- SAS Hadoop Configuration Guide for Base SAS and SAS/ACCESS
- SAS LASR Analytic Server: Reference Guide

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

SAS Books
SAS Campus Drive
Cary, NC 27513-2414
Phone: 1-800-727-0025
Fax: 1-919-677-4444
Email: sasbook@sas.com
Web address: sas.com/store/books
<table>
<thead>
<tr>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>C</strong></td>
</tr>
<tr>
<td>co-located deployment</td>
</tr>
<tr>
<td><strong>D</strong></td>
</tr>
<tr>
<td>deployment approaches</td>
</tr>
<tr>
<td>for data management</td>
</tr>
<tr>
<td>for in-memory analytics</td>
</tr>
<tr>
<td><strong>H</strong></td>
</tr>
<tr>
<td>Hadoop Distributed File System (HDFS)</td>
</tr>
<tr>
<td>Hadoop distributions</td>
</tr>
<tr>
<td>configuring for in-memory analytics</td>
</tr>
<tr>
<td>SAS High-Performance Deployment of Hadoop</td>
</tr>
<tr>
<td>supported</td>
</tr>
<tr>
<td>HCatalog</td>
</tr>
<tr>
<td>HDFS (Hadoop Distributed File System)</td>
</tr>
<tr>
<td>Hive</td>
</tr>
<tr>
<td>HiveServer2</td>
</tr>
<tr>
<td><strong>I</strong></td>
</tr>
<tr>
<td>in-memory analytics</td>
</tr>
<tr>
<td>co-located deployment</td>
</tr>
<tr>
<td>deploy on a Hadoop cluster</td>
</tr>
<tr>
<td>deploy with SAS High-Performance Deployment of Hadoop</td>
</tr>
<tr>
<td>explained</td>
</tr>
<tr>
<td>reasons to deploy</td>
</tr>
<tr>
<td>related products</td>
</tr>
<tr>
<td>remote access to Hadoop</td>
</tr>
<tr>
<td>software components</td>
</tr>
<tr>
<td><strong>K</strong></td>
</tr>
<tr>
<td>Kerberos</td>
</tr>
<tr>
<td><strong>M</strong></td>
</tr>
<tr>
<td>MapReduce</td>
</tr>
<tr>
<td><strong>O</strong></td>
</tr>
<tr>
<td>Oozie</td>
</tr>
</tbody>
</table>
passwordless SSH 21, 25
processing data
  basic access 2
  in a Hadoop cluster 1
processing data, in memory
  See in-memory analytics

QKB (SAS Quality Knowledge Base) 8

SAS Code Accelerator for Hadoop 2
SAS Data Loader for Hadoop deployment steps 9
directives 7, 10
der user configuration 10
explained 2, 7
software components 8
vApp 8
SAS Data Quality Accelerator for Hadoop 8
SAS Deployment Wizard 4
SAS Embedded Process explained 1
purpose, data management 8
purpose, in-memory analytics 16
remote access to Hadoop 28
SAS High-Performance Analytics environment 2, 15, 19, 22, 23, 26
SAS High-Performance Analytics node installation 15
SAS High-Performance Analytics products 18
SAS High-Performance Computing Management Console 16, 19, 21, 23, 25
SAS High-Performance Deployment of Hadoop 19, 21
SAS In-Database Code Accelerator for Hadoop 8
SAS in-database deployment package for Hadoop
  See SAS Embedded Process
SAS In-Memory Statistics 18
SAS Intelligence Platform 4
SAS LASR Analytic Server 2, 14, 15, 21, 25
SAS Quality Knowledge Base (QKB) 8
SAS Scoring Accelerator for Hadoop 2
SAS Software Depot 4
SAS Visual Analytics 2, 18
SAS Visual Statistics 2
SAS/ACCESS Interface to Hadoop explained 1, 8
purpose, in-memory analytics 16
remote access to Hadoop 28
SASHDAT tables 13, 17
Sqoop 3

TKGrid 15, 22, 26

TKGrid_REP 29

Y

YARN 3, 17
Index
Gain Greater Insight into Your SAS® Software with SAS Books.

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

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