THE POWER TO KNOW.

SAS® and Hadoop Technology: Overview

SAS® Documentation
August 24, 2020

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2015. *SAS® and Hadoop Technology: Overview*. Cary, NC: SAS Institute Inc.

SAS® and Hadoop Technology: Overview

ISBN 978-1-62959-983-0 (Paperback)
ISBN 978-1-62959-986-1 (Epub)
ISBN 978-1-62959-987-8 (Mobi)
ISBN 978-1-62959-984-7 (PDF)

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

August 2020

9.4-P4:hadoopov

Contents

Chapter 1
Introduction to SAS and Hadoop Technology

SAS and Hadoop—Natural Complements

When you use SAS with Hadoop, you combine the power of analytics with the key strengths of Hadoop: large-scale data processing and commodity-based storage and compute resources. Using SAS with Hadoop maximizes big data assets in the following ways:

- Hadoop data can be leveraged using SAS. Just as with other data sources, data stored in Hadoop can be transparently consumed by SAS. This means that tools already in place can be used with Hadoop. Not only can SAS access data from Hadoop, but SAS can also assist in managing your Hadoop data.
- The power of SAS analytics is extended to Hadoop. Long before the term big data was coined, SAS applied complex analytical processes to large volumes of data. SAS was designed from the beginning to scale and perform well in any environment and to take advantage of complementary technologies.

Currently, more than 20 SAS products, solutions, and technology packages interact with Hadoop. Each SAS technology provides different functionality—from accessing and managing Hadoop data to executing analytical models in a Hadoop cluster. In addition to the variety of functionality, SAS processes Hadoop data using different methods so that a particular business problem can be resolved in an optimal way.

About This Document

This document provides an overview of SAS and Hadoop technology and explains how SAS and Hadoop work together. Use this document as a starting point to learn about the SAS technology that interacts with Hadoop. No matter how much you know about SAS or Hadoop, getting acquainted with the concepts enables you to understand and use the technology that best meets your specific needs.

The information in this document is useful for the following audience:

- IT administrators who are interested in what SAS can do with Hadoop
- SAS customers who are considering moving their data to Hadoop and who want to know how their SAS products interact with Hadoop
- prospective SAS customers who want to know whether SAS and Hadoop technology can resolve their business problems

The following information is provided:

- Chapter 2, “Why Hadoop?,” on page 3 introduces Hadoop concepts, such as what Hadoop is, benefits of storing data in Hadoop, Hadoop components, Hadoop distributions, and basic information about connecting SAS to Hadoop.
- Chapter 3, “How Do SAS and Hadoop Work Together?,” on page 9 provides concepts about how SAS processes Hadoop data by eliminating or reducing data movement. In addition, there is information about deploying and securing the SAS and Hadoop environment.
- Chapter 4, “What SAS Technology Interacts with Hadoop?,” on page 17 introduces SAS technology that interacts with Hadoop. Examples of what you can do with SAS and Hadoop are provided, and each SAS technology is listed by its function with a description.
- Chapter 5, “Explore Data and Develop Models,” on page 21 provides a summary of each SAS technology that explores and visualizes data and develops analytical models. These technologies include SAS Visual Analytics, SAS Visual Statistics, SAS In-Memory Statistics, SAS High-Performance Analytics products, SAS High-Performance Risk, and SAS In-Database Technology.
- Chapter 6, “Execute Models,” on page 31 provides a summary of the SAS Scoring Accelerator for Hadoop, which executes analytical models in a Hadoop cluster.
- Chapter 7, “Manage Data,” on page 33 provides a summary of each SAS technology that accesses and manages data. These technologies include SAS Data Loader for Hadoop, accelerators that enable SAS code to be executed in a Hadoop cluster, several SAS/ACCESS engines, Base SAS functionality, and SAS Data Integration Studio.
- Chapter 8, “Additional Functionality,” on page 49 provides a summary of additional SAS functionality such as SAS Event Stream Processing, SAS Federation Server, SAS Grid Manager for Hadoop, SAS Scalable Performance Data (SPD) Server, and SAS Visual Scenario Designer.

Resources for SAS Viya

This guide provides information about SAS 9.4 and Hadoop. To find information about SAS Viya, see the following resources:

- SAS Viya product page
- SAS Viya documentation page

Chapter 2
Why Hadoop?

What Is Apache Hadoop?

Apache Hadoop is an open-source software framework that provides massive data storage and distributed processing of large amounts of data. The Hadoop framework provides the tools needed to develop and run software applications.

Data is divided into blocks and stored across multiple connected nodes (computers) that work together. This setup is referred to as a cluster. A Hadoop cluster can span thousands of nodes. Computations are run in parallel across the cluster, which means that the work is divided among the nodes in the cluster.

Hadoop runs on a Linux operating system. Hadoop is available from either the Apache Software Foundation or from vendors that offer their own commercial Hadoop distributions such as Cloudera, Hortonworks, IBM BigInsights, MapR, and Pivotal.

For more information about Hadoop, see Welcome to Apache Hadoop.

Benefits of Storing Data in Hadoop

The benefits of storing data in Hadoop include the following:

- Hadoop accomplishes two tasks: massive data storage and distributed processing.
- Hadoop is a low-cost alternative for data storage over traditional data storage options. Hadoop uses commodity hardware to reliably store large quantities of data.
- Data and application processing are protected against hardware failure. If a node goes down, data is not lost because a minimum of three copies of the data exist in the Hadoop cluster. Furthermore, jobs are automatically redirected to working machines in the cluster.
- The distributed Hadoop model is designed to easily and economically scale up from single servers to thousands of nodes, each offering local computation and storage.
- Unlike traditional relational databases, you do not have to preprocess data before storing it in Hadoop. You can easily store unstructured data.
- You can use Hadoop to stage large amounts of raw data for subsequent loading into an enterprise data warehouse or to create an analytical store for high-value activities such as advanced analytics, querying, and reporting.

Hadoop Platform

Overview of Hadoop Platform

Hadoop consists of a family of related components that are referred to as the Hadoop ecosystem. Hadoop provides many components such as the core components HDFS (Hadoop Distributed File System) and MapReduce. In addition, Hadoop software and services providers (such as Cloudera and Hortonworks) provide additional proprietary software.

Hadoop Components and Other Related Components

Ambari and Cloudera Manager

Ambari is an open-source, web-based tool for managing, configuring, and testing Hadoop services and components. Ambari can be used with Hortonworks, IBM BigInsights, and Pivotal. Cloudera Manager is a cluster manager tool that can be used with Cloudera.

HBase

HBase is an open-source, non-relational, distributed database that runs on top of HDFS. HBase tables can serve as input for and output of MapReduce programs.

HDFS

HDFS provides distributed data storage and processing. HDFS is fault-tolerant, scalable, and simple to expand. HDFS manages files as blocks of equal size, which are replicated across the machines in a Hadoop cluster. HDFS stores all types of data without prior organization such as Microsoft Excel spreadsheets, Microsoft Word documents, videos, and so on. HDFS supports all types of data formats. MapReduce is used to read the different formats.

HDFS includes various shell-like commands for direct interaction. These commands support most of the normal file system operations like copying files and changing file permissions, as well as advanced operations such as setting file redundancy to a different replication number.

Hive and HiveServer2

Hive is a distributed data warehouse component that is built on top of HDFS. The original Hive was succeeded by HiveServer2. The terms “Hive” and “HiveServer2” have become interchangeable, but mostly refer to HiveServer2. Hive provides the SQL query language HiveQL for data queries, analysis, and summarization. HiveServer2 can be secured with the Lightweight Directory Access Protocol (LDAP), which is a directory service protocol that authenticates users to a computer system. Or, it can be secured with Kerberos, which is a network authentication protocol that enables nodes to verify their identities to one another using tickets.

HiveQL

HiveQL is the SQL query language for Hive and HiveServer2.

Oozie

Oozie is a workflow scheduler system that manages Hadoop jobs.

MapReduce

MapReduce is a parallel programming model that is built into Hadoop for distributed processing. MapReduce divides applications into smaller components and distributes them among numerous machines. The map phase performs operations such as filtering, transforming, and sorting. The reduce phase takes the output and aggregates it. The second generation of MapReduce is referred to as YARN (Yet Another Resource Negotiator).

Pig

Pig is a platform for analyzing very large data sets that are stored in HDFS. Pig consists of a compiler for MapReduce programs and a high-level language called Pig Latin. Pig Latin provides a way to perform data extractions, transformations, loading, and basic analysis without having to write MapReduce programs.

Spark

Spark is an open-source, distributed processing system used for big data workloads. It utilizes in-memory caching and optimized query execution for fast queries against data of any size. SAS issues SparkSQL queries to access structured data on a Spark cluster.

Sqoop

Sqoop is open-source software that transfers data between a relational database and Hadoop.

YARN

YARN is a resource-management platform for scheduling and handling resource requests from a distributed application. YARN refers to the second generation of MapReduce.

ZooKeeper

ZooKeeper is open-source software that provides coordination services for distributed applications. It exposes common services (such as naming, configuration management, and synchronization) and group services.

HAWQ

HAWQ (Hadoop With Query) is an SQL engine that is provided by Pivotal. HAWQ provides an optimized Hadoop SQL query mechanism on top of Hadoop. HAWQ provides ANSI SQL support and enables SQL queries of HBase tables. HAWQ includes a set of catalog services and does not use the Hive metastore.

Impala

Impala is an open-source massively parallel processing query engine that is provided by Cloudera and MapR. You use Impala to issue HiveQL queries to data stored in HDFS and HBase without moving or transforming data.

Kerberos

Kerberos is an open-source computer network authentication protocol that enables nodes to verify their identities to one another using tickets. Kerberos was developed as part of the Athena Project at the Massachusetts Institute of Technology (MIT). The Kerberos protocol is implemented as a series of negotiations between a client, the authentication server, and the service server. Secure authentication of Hadoop clusters has been available using the Kerberos protocol since Hadoop 2.

Sentry

Sentry is an open-source authorization mechanism that provides fine-grained and role-based access control for Apache Hive and Cloudera Impala. Sentry is a fully integrated component of CDH, which is a Cloudera distribution of Hadoop and related projects.

RecordService

RecordService is an open-source security layer that enforces fine-grained access control to data.

Hadoop Distributions

Hadoop is available from the following sources:

- Apache Software Foundation
- Commercial Hadoop distributions

A commercial Hadoop distribution is the collection of Hadoop components (such as HDFS, Hive, and MapReduce) that is provided by a vendor. Many commercial Hadoop distributions include additional proprietary software. SAS supports commercial Hadoop distributions from Cloudera, Hortonworks, IBM BigInsights, MapR, and Pivotal.

TIP Each SAS technology does not support all commercial Hadoop distributions. For more information about supported commercial Hadoop distributions, see the website SAS 9.4 Support for Hadoop.

Connecting to Hadoop

Hadoop Cluster Configuration Files

Hadoop cluster configuration files are key to communicating with the Hadoop cluster. The configuration files define how to connect to the Hadoop cluster and they provide other system information. The default Hadoop configuration consists of two types of configuration files: default files and site-specific files. The site-specific configuration files include multiple files, such as core-site.xml, hdfs-site.xml, hive-site.xml, mapred-site.xml, and yarn-site.xml.

TIP Some SAS technology requires that you perform steps to make the Hadoop cluster configuration files accessible to the SAS client machine. See the SAS technology summaries in this document to determine what is required.

Hadoop Distribution JAR Files

JAR files are compressed collections of Java class files. The Hadoop distribution JAR files contain the Java application code deployed to the SAS client machine to enable SAS to connect to Hadoop as a client. JAR files are similar to client tools for relational databases.

The JAR files are specific to the version of the Hadoop distribution and the Hadoop components that you are using. That is, if you update your Hadoop environment, the JAR files need to be updated on the SAS client as well. In many cases, if SAS code is failing, it is because of a missing JAR file or a JAR file version mismatch between the SAS client and server.

TIP Some SAS technology requires that you perform steps to make the Hadoop distribution JAR files accessible to the SAS client machine. See the SAS technology summaries in this document to determine what is required. In addition, some SAS technology, such as the SAS Embedded Process, requires installation of JAR files that are provided by SAS.

SAS Plug-ins for Hadoop

SAS Plug-ins for Hadoop is a group of JAR files and executables that enable SAS LASR Analytic Server and other applications that rely on the SAS High-Performance Analytics environment to work with SASHDAT files in a co-located environment.

HttpFS

HttpFS is a server that provides a REST HTTP gateway supporting all HDFS operations.

TIP SAS technology that connects to HDFS using HttpFS requires that you perform specific steps to connect. See the SAS technology summaries in this document to determine what is required.

WebHDFS

WebHDFS is an HTTP REST API that supports the complete file system interface for HDFS.

TIP SAS technology that connects to HDFS using WebHDFS requires that you perform specific steps to connect. See the SAS technology summaries in this document to determine what is required.

Chapter 3

How Do SAS and Hadoop Work Together?

Understanding Data Movement

Overview

For a computer program to change data into information, there are three basic steps:

1. Access the data (which is stored in a data source).
2. Process the data.
3. Use the results somewhere (such as, write a report).

Moving data to be processed can take a lot of time, especially big data. If you can limit data movement or improve getting data to and from processing, then you can provide results faster.

Traditional SAS processing involves extracting data from the data source and delivering it to the SAS server for processing. Today, in addition to traditional processing, SAS uses alternative methods to work with Hadoop so that data movement can be eliminated or reduced. As a result, a particular business problem can be resolved in an optimal way.

To understand the SAS technology that interacts with Hadoop, it is helpful to understand how SAS works with Hadoop. The following topics describe the different methods that SAS uses, starting with processing directly in the Hadoop cluster, moving the processing closer to the data in a SAS in-memory environment, and performing traditional processing.

Processing in the Hadoop Cluster

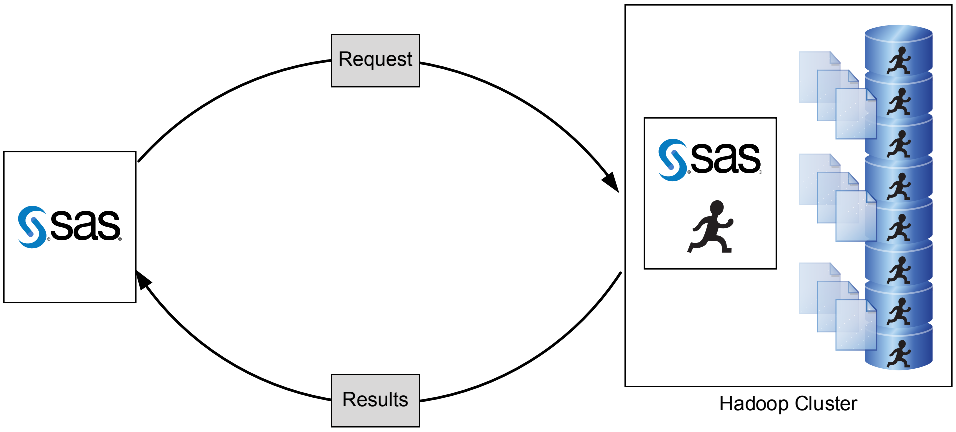
To eliminate data movement, SAS can process data directly in the Hadoop cluster. To do this, SAS pushes the SAS code directly to the nodes of the Hadoop cluster for processing. Rather than extracting the data from a data source and delivering it to a SAS server, SAS brings the analytics to where the data is stored to take advantage of the distributed processing capabilities of Hadoop.

The SAS technology that processes data in the Hadoop cluster provides the following advantages:

- SAS computations are orchestrated via the distributed processing capabilities of Hadoop, which translates to shorter processing times and more value from Hadoop itself.
- Data movement is reduced and data security is improved.
- All of the data can be used in calculations versus a sample of the data, thus you gain accuracy in results.
- Processing directly in the Hadoop cluster is advantageous for mature Hadoop environments when data is so voluminous that moving it is prohibitive.

In the following illustration, the SAS server or SAS client connects to the Hadoop cluster, submits a request, processes the request in the Hadoop cluster, and sends only the results back to SAS.

Figure 3.1 Processing in the Hadoop Cluster



TIP SAS Data Loader and SAS In-Database Technology can process data in the Hadoop cluster. In addition, SAS/ACCESS Interface to Hadoop can pass SQL code to the Hadoop cluster, the SAS Scalable Performance Data (SPD) Engine can submit data subsetting to the Hadoop cluster, and PROC HADOOP enables you to submit MapReduce programs and Pig Latin code for further processing by Hadoop.

Processing in a SAS In-Memory Environment

Some SAS technology works with Hadoop by processing data in a SAS in-memory environment. The in-memory environment exists on an analytics cluster, which is a set of connected machines with in-memory SAS software. The in-memory SAS software consists of the SAS High-Performance Analytics environment and the SAS LASR Analytic Server, which you will learn more about in "SAS Technology That Minimizes Data Movement" on page 13.

To process the data in an in-memory environment, SAS loads the data from Hadoop into the in-memory environment. The in-memory environment performs the analysis, and only the results are returned to the SAS server or SAS client that submitted the request.

SAS technology that processes Hadoop data in an in-memory environment provides the following advantages:

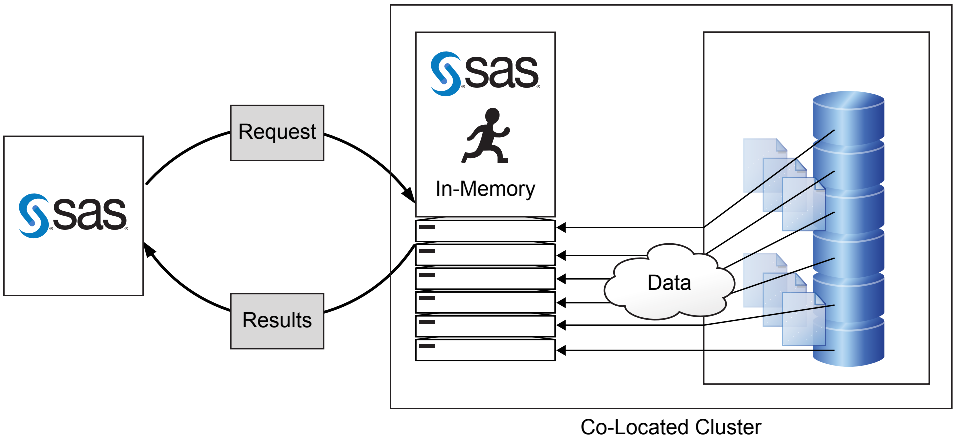
- Data is loaded into the in-memory environment in parallel, which avoids the network bandwidth limitations of a single network connection.
- The in-memory data is distributed among multiple machines and treated as one large object, which provides fast results.
- SAS keeps the data and computations massively parallel.
- The in-memory environment brings the analytics closer to the data, which reduces time-consuming data movement.

The in-memory environment can be configured in the following ways:

- on the Hadoop cluster that has the data to analyze, referred to as "co-located"
- on a set of machines that is remote from the Hadoop cluster and dedicated to SAS processing

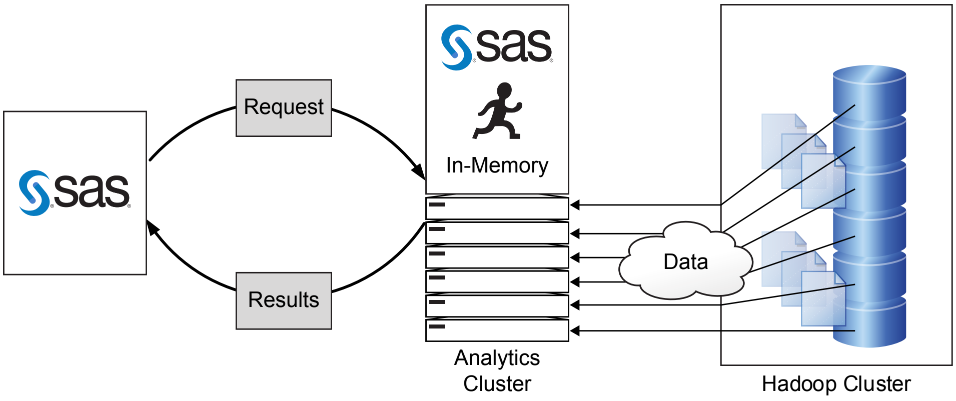
In the following illustration, the in-memory environment is co-located on the Hadoop cluster. The SAS server or SAS client connects to the Hadoop cluster, submits a request, loads the Hadoop data into the in-memory environment, processes the request, and sends only the results back to SAS.

Figure 3.2 *Processing in an In-Memory Environment That Is Co-Located*



In the following illustration, the in-memory environment is on a separate set of machines from the Hadoop cluster. The SAS server or client connects to the analytics cluster that is remote from the Hadoop cluster, submits a request, loads the Hadoop data to the in-memory environment, processes the request, and then sends only the results back to SAS.

Figure 3.3 *Processing in an In-Memory Environment That Is Remote from the Hadoop Cluster*



TIP SAS Visual Analytics, SAS In-Memory Statistics, SAS High-Performance Analytics products (such as SAS High-Performance Data Mining, SAS High-Performance Econometrics, SAS High-Performance Optimization, SAS High-Performance Statistics, and SAS High-Performance Text Mining), SAS High-Performance Risk, and SAS Visual Scenario Designer can process Hadoop data in an in-memory environment.

Traditional Processing

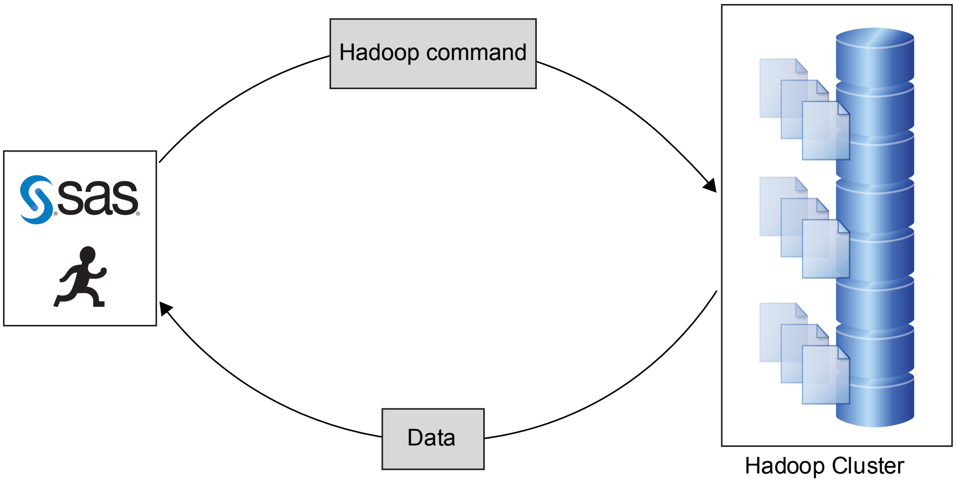
For traditional SAS environments, a SAS server runs on a single machine, reads data from files or network connections, and processes the data locally on the machine. Applying this traditional model to Hadoop, SAS provides a bridge to Hadoop to move data to and from Hadoop. SAS connects to the Hadoop cluster, extracts the data, and delivers it to the SAS server for processing.

SAS technology that accesses and extracts data from Hadoop provides the following advantages:

- To SAS, Hadoop is simply another data source like a SAS data set or a third-party database.
- You can manage and analyze Hadoop data with any of your favorite SAS tools.

In the following illustration, the SAS server connects to the Hadoop cluster, submits a request, extracts the Hadoop data, delivers it to the SAS server, and then processes the request.

Figure 3.4 Accessing and Extracting Hadoop Data



TIP SAS/ACCESS Interface to Hadoop, Base SAS FILENAME statement, Hadoop access method, and SPD Engine can access and extract data from Hadoop.

SAS Technology That Minimizes Data Movement

Overview

SAS uses several software components to interact with Hadoop to reduce or eliminate data movement. For some SAS technology, the SAS High-Performance Analytics environment and the SAS LASR Analytic Server provide in-memory environments. To push processing directly in the Hadoop cluster, some SAS technology uses the SAS Embedded Process and SAS accelerators.

SAS Embedded Process

The SAS Embedded Process is a software component that is installed and runs on the Hadoop cluster. The SAS Embedded Process is the core technology that supports the following functionality:

- To process a request in the Hadoop cluster, the SAS Embedded Process and SAS/ACCESS Interface to Hadoop work with the SAS Contextual Analysis In-Database Scoring for Hadoop, SAS In-Database Code Accelerator for Hadoop, SAS Data Quality Accelerator for Hadoop, and SAS Scoring Accelerator for Hadoop to push processing directly in the Hadoop cluster to read and write data in parallel.
- To process a request in an in-memory environment, the SAS Embedded Process provides a high-speed parallel connection that loads data from Hadoop to the SAS High-Performance Analytics environment and the SAS LASR Analytic Server.

Basically, the SAS Embedded Process is a subset of Base SAS software that is sufficient to support the multithreaded SAS DS2 language. The SAS Embedded Process runs in its own processing space in Hadoop. Each node of the Hadoop cluster runs one instance of the SAS Embedded Process. Each instance serves all of the threads of query parallelism executing on that node at a given time. On a Hadoop cluster, a special set of MapReduce classes associates the SAS Embedded Process with each task.

TIP SAS technology that processes Hadoop data directly in the Hadoop cluster or in an in-memory environment might require that the SAS Embedded Process be installed on the Hadoop cluster. See the SAS technology summaries in this document to determine what is required.

SAS High-Performance Analytics Environment

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

With the SAS High-Performance Analytics environment, operations are processed in a scalable, in-memory environment. In this environment, the data is loaded into memory, the analysis is performed, and then the in-memory resources are freed.

TIP The SAS High-Performance Analytics products (such as SAS High-Performance Data Mining, SAS High-Performance Econometrics, SAS High-Performance Optimization, SAS High-Performance Statistics, and SAS High-Performance Text Mining) and SAS High-Performance Risk use the SAS High-Performance Analytics environment.

SAS LASR Analytic Server

The SAS LASR Analytic Server is a scalable, analytic platform that provides a secure, multi-user environment for concurrent access to in-memory data. SAS LASR Analytic Server provides the ability to load Hadoop data into memory and perform distributed processing, exploratory analysis, analytic calculations, and more—all interactively.

Once data is loaded into memory, it remains in memory for simultaneous access by any number of users until the data is explicitly unloaded from memory. In-memory persistence avoids unnecessary and expensive multiple data loading steps. By reading the data into memory only once, it provides fast interactive ad hoc analysis and data management, resulting in greater productivity.

The SAS LASR Analytic Server supports HDFS as a co-located cluster deployment, which means that SAS and Hadoop are installed on the same set of machines. For this processing, SAS uses the memory resources of the set of machines as a computational space rather than as a database.

The SAS LASR Analytic Server provides the following components:

LASR procedure
administers the SAS LASR Analytic Server and enables loading data in parallel from HDFS. When combined with SAS/ACCESS Interface to Hadoop and the SAS Embedded Process, the LASR procedure can load data in parallel in formats other than SASHDAT.

SAS LASR Analytic Server engine (also called the SASIOLA engine)
loads data to memory serially. The engine loads data from a SAS data set or from any data source that SAS can access when the SAS data set is small or parallel loading is not possible.

SASHDAT engine (previously called SAS Data in HDFS engine)
adds and deletes SASHDAT files. A SASHDAT file is in a SAS proprietary file format that is designed for high performance to load and unload into memory very fast. The SASHDAT engine enables SAS LASR Analytic Server and high-performance procedures to read CSV files.

TIP SAS Visual Analytics, SAS Visual Statistics, SAS In-Memory Statistics, SAS Data Loader, and SAS Visual Scenario Designer use the SAS LASR Analytic Server.

Deploying the SAS and Hadoop Environment

Because much of high-performance analytics is designed to run with a distributed processing system like Hadoop, SAS analytics require a hardware environment so that computations are run in parallel. The key is a set of multiple connected computers that work together, which is often a system of servers that is referred to as an "analytics cluster." Computations are run in parallel across the analytics cluster, which means that the work is divided among the nodes in the cluster.

To deploy SAS and Hadoop, you can do the following:

- co-locate SAS software with Hadoop. That is, SAS and Hadoop exist on the same set of machines.
- have SAS software remote from the Hadoop cluster. That is, one set of machines is dedicated to SAS software, and the Hadoop cluster exists on a separate set of machines.

Each SAS solution or product has documentation that helps you deploy the software and information that helps you configure SAS software with Hadoop. Because some SAS deployments require multiple SAS solutions, products, and additional software, see *SAS and Hadoop Technology: Deployment Scenarios* for deployment examples.

Securing the SAS and Hadoop Environment

Kerberos

The SAS technology that interacts with Hadoop supports the Kerberos authentication protocol.

To have a fully operational and secure Hadoop environment, it is critical to understand the requirements and preparation for and process around Kerberos enablement. There are four overall practices that help ensure that your SAS and Hadoop connection is secure and that SAS performs well within the environment.

1. Understand the fundamentals of Kerberos authentication and the best practices promoted by Hadoop providers.
2. Simplify Kerberos setup by installing SAS and Hadoop on the same set of machines.
3. Ensure that Kerberos prerequisites are met when installing and configuring SAS applications that interact with Hadoop.
4. When configuring SAS and Hadoop jointly in a high-performance environment, ensure that all SAS servers are recognized by Kerberos.

Secure data and secure user authentication are critical requirements for enterprise implementations of Hadoop. For details about the supported Hadoop security configurations, see the website Support for Hadoop.

Chapter 4

What SAS Technology Interacts with Hadoop?

Spanning the Data-to-Decision Life Cycle

SAS offers technology that interacts with Hadoop to bring the power of SAS analytics to Hadoop and spans the entire data-to-decision life cycle. Using SAS technology that interacts with Hadoop, you can do the following:

- access and manage your Hadoop data
- explore data and develop models
- execute analytical models in Hadoop

Here are a few examples of what you can do with SAS technology that interacts with Hadoop:

- With SAS/ACCESS Interface to Hadoop, you can connect to a Hadoop cluster and read and write data to and from Hadoop. You can analyze Hadoop data with your favorite SAS procedures and the DATA step.
- Suppose you want to connect to Hadoop, read and write data, or execute a MapReduce program. Using Base SAS, you can simply use the FILENAME statement with the Hadoop access method to read data from HDFS and write data to HDFS. You can use the HADOOP procedure to submit HDFS commands, MapReduce programs, and Pig Latin code. For example, you could use PROC HADOOP to create a directory in HDFS, and then use the FILENAME statement to copy a SAS data set to the new HDFS directory.
- SAS/ACCESS Interface to Impala provides direct, transparent access to Cloudera Impala and MapR Impala from your SAS session.
- SAS/ACCESS Interface to HAWQ provides access to Hadoop data via the HAWQ engine.
- SAS/ACCESS Interface to Spark provides direct, transparent access to Spark data from your SAS session, using SparkSQL without requiring familiar with DataFrame API.
- The SPD Engine enables you to interact with Hadoop through HDFS. You can write data, retrieve data for analysis, perform administrative functions, and even update

data as an SPD Engine data set. The SPD Engine organizes data into a streamlined file format that has advantages for a distributed file system like HDFS.

- With SAS Data Loader for Hadoop, you can copy data to and from Hadoop. In addition, you can profile, cleanse, query, transform, and analyze data in Hadoop.
- With SAS Visual Analytics, you can explore and visualize large amounts of data stored in HDFS, and then create and modify predictive models using a visual interface and in-memory processing. In addition, you can publish reports to the web and mobile devices.
- SAS High-Performance Analytics products provide a highly scalable in-memory infrastructure that supports Hadoop. SAS provides high-performance procedures that enable you to manipulate, transform, explore, model, and score data all within Hadoop.
- Using SAS In-Database Technology, certain SAS procedures, DATA step programs, data quality operations, DS2 threaded programs, and scoring models can be submitted and executed in Hadoop. In-database processing uses the distributed processing capabilities of Hadoop to process the requests.
- Using SAS In-Memory Statistics, you can work with your Hadoop data to perform analytical data preparation, variable transformations, exploratory analysis, statistical modeling and machine-learning techniques, integrated modeling comparison, and model scoring.

SAS Technology That Interacts with Hadoop

The following table lists each SAS technology that interacts with Hadoop, its function, and its description. See each SAS technology for a summary that provides a description, features, what is required to execute the software, and references to the full product documentation.

Table 4.1 *SAS Technology That Interacts with Hadoop*

Function	SAS Technology	Description
Explore Data and Develop Models	"SAS Visual Analytics and SAS Visual Statistics" on page 21	Explores and visualizes huge volumes of data to identify patterns and trends and opportunities for further analysis.
	"SAS In-Memory Statistics" on page 23	Performs analytical data preparation, variable transformations, exploratory analysis, statistical modeling and machine-learning techniques, integrated modeling comparison, and model scoring.

Function	SAS Technology	Description
	“SAS High-Performance Analytics Products” on page 25	Provides tools that perform analytic tasks in a high-performance environment to provide data mining, text mining, econometrics, and optimization capabilities.
	“SAS High-Performance Risk” on page 28	Provides a financial portfolio management solution that enables you to price very large portfolios for thousands of market states.
	“SAS In-Database Technology” on page 29	Executes select SAS processing in Hadoop, such as in-database SAS procedures, DATA step programs, DS2 threaded programs, and scoring models.
Execute Models	“SAS Scoring Accelerator for Hadoop” on page 31	Executes analytical models in a Hadoop cluster.
Manage Data	“SAS Data Loader for Hadoop” on page 34	Transforms, queries, profiles, and analyzes big data without moving the data.
	“SAS Data Quality Accelerator for Hadoop” on page 35	Provides in-database data quality operations in a Hadoop cluster.
	“SAS In-Database Code Accelerator for Hadoop” on page 36	Executes DS2 code in a Hadoop cluster.
	“SAS/ACCESS Interface to Hadoop” on page 36	Accesses Hadoop data through HiveServer2 and from HDFS.
	“SAS/ACCESS Interface to HAWQ” on page 38	Accesses the Pivotal HAWQ SQL engine.
	“SAS/ACCESS Interface to Impala” on page 39	Accesses Impala.
	“SAS/ACCESS Interface to Spark” on page 40	Accesses data on a Spark cluster through SparkSQL.
	“SQOOP Procedure” on page 41	Accesses Apache Sqoop to transfer data between a database and HDFS.

Function	SAS Technology	Description
	"Base SAS FILENAME Statement with the Hadoop Access Method" on page 42	Reads data from and writes data to HDFS using the SAS DATA step.
	"Base SAS HADOOP Procedure" on page 43	Submits HDFS commands, MapReduce programs, and Pig Latin code from your SAS session.
	"SAS Scalable Performance Data (SPD) Engine" on page 44	Interacts with Hadoop through HDFS to write data, retrieve data for analysis, perform administrative functions, and update data as an SPD data set.
	"SAS Data Integration Studio" on page 46	Builds, implements, and manages data integration processes.
Additional Functionality	"SAS Event Stream Processing" on page 49	Builds applications that can process and analyze volumes of continuously flowing event streams.
	"SAS Federation Server" on page 50	Provides scalable, threaded, multi-user, and standards-based data access technology by integrating data from multiple data sources.
	"SAS Grid Manager for Hadoop" on page 52	Provides workload management, accelerated processing, and scheduling of SAS analytics on your Hadoop cluster.
	"SAS Scalable Performance Data (SPD) Server" on page 52	Provides a multi-user, high-performance data delivery environment that enables you to interact with Hadoop through HDFS.
	"SAS Visual Scenario Designer" on page 54	Identifies events or patterns that might be associated with fraud or non-compliance.

Chapter 5

Explore Data and Develop Models

SAS Visual Analytics and SAS Visual Statistics

What Is SAS Visual Analytics?

SAS Visual Analytics leverages SAS high-performance analytic technologies and empowers organizations to explore huge volumes of data quickly to identify patterns, trends, and opportunities for further analysis. The highly visual, drag-and-drop data interface of SAS Visual Analytics, combined with the speed of SAS LASR Analytic Server, accelerates analytic computations and enables organizations to derive value from massive amounts of data. This creates an unprecedented ability to solve difficult problems, improve business performance, predict future performance, and mitigate risk rapidly and confidently. Users can quickly design reports or dashboards, which can be viewed on a mobile device or on the web.

What Is SAS Visual Statistics?

SAS Visual Statistics is an add-on to SAS Visual Analytics that enables you to develop and test models using the in-memory capabilities of SAS LASR Analytic Server. SAS Visual Analytics enables you to explore, investigate, and visualize data sources to uncover relevant patterns. SAS Visual Statistics extends these capabilities by creating, testing, and comparing models based on the patterns discovered in SAS Visual Analytics. SAS Visual Statistics can export the score code, before or after performing model comparison, for use with other SAS products and to put the model into production.

Why Use SAS Visual Analytics and SAS Visual Statistics?

Using SAS Visual Analytics, you can explore new data sources, investigate them, and create visualizations to uncover relevant patterns. You can then easily share those visualizations in reports. In traditional reporting, the resulting output is well-defined up-front. That is, you know what you are looking at and what you need to convey. However, data discovery requires that you understand the data, its characteristics, and its relationships. Then, when useful visualizations are created, you can incorporate those visualizations into reports that are available on a mobile device or in the viewer.

SAS Visual Analytics provides the following benefits:

- enables users to apply the power of SAS analytics to massive amounts of data
- empowers users to visually explore data, based on any variety of measures, at amazingly fast speeds
- enables users to quickly create reports or dashboards using standard tables, graphs, and gauges
- enables users to share insights with anyone, anywhere, via the web or a mobile device

SAS Visual Statistics provides the following benefits:

- enables users to rapidly create powerful predictive and descriptive models using all data and the latest algorithms in an easy-to-use, web-based interface
- enables users to compare the relative performance of two or more competing models using a variety of criteria to choose a champion model
- enables users to export model score code for any model so that users can easily apply the model to new data and get timely results

What Is Required?

- You must license SAS Visual Analytics. The package includes a restricted version of Base SAS 9.4 and the SAS LASR Analytic Server.
- SAS Visual Statistics is an add-on to SAS Visual Analytics.
- To load data into memory, you can use the LASR procedure to load data in parallel or the SAS LASR Analytic Server engine to load data serially. Or, if your Hadoop

cluster has data in Hive or HiveServer2, you can use SAS/ACCESS Interface to Hadoop to load the data into memory.

- If the SAS LASR Analytic Server is co-located with the Hadoop cluster, SASHDAT and CSV files are automatically loaded in parallel. To load other data in parallel, the SAS Embedded Process must be installed on the Hadoop cluster.
- If the SAS LASR Analytic Server is installed remotely from the Hadoop cluster, to load any data in parallel, the SAS Embedded Process must be installed on the Hadoop cluster. In addition, you must license SAS/ACCESS Interface to Hadoop.
- SAS/ACCESS Interface to Hadoop resides on the SAS client machine that you use for submitting SAS programs.
- SAS Visual Analytics requires the SAS Intelligence Platform. The system administrator must install and configure the required SAS Intelligence Platform software. In addition, the system administrator must use SAS Management Console to maintain metadata for servers, users, and other global resources that are required by SAS Visual Analytics.

More Information

- For more information about how to use SAS Visual Analytics and SAS Visual Statistics, see the Documentation section of the SAS Visual Analytics support page.
- For administration of SAS Visual Analytics and SAS Visual Statistics, see *SAS Visual Analytics: Administration Guide*.
- For more information about the SAS LASR Analytic Server, including the LASR procedure, SASIOLA engine, and SASHDAT engine, see *SAS LASR Analytic Server: Reference Guide*.
- For SAS LASR Analytic Server deployment, installation, and configuration information, see *SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide*.
- For information about installing the SAS Embedded Process, see *SAS In-Database Products: Administrator's Guide*.

SAS In-Memory Statistics

What Is SAS In-Memory Statistics?

SAS In-Memory Statistics provides the data scientist or analytical expert with interactive programming access to large data sets stored in Hadoop. With SAS In-Memory Statistics, you load data from Hadoop into a SAS in-memory environment, and then perform analytical data preparation, variable transformations, exploratory analysis, statistical modeling and machine-learning techniques, integrated modeling comparison, and model scoring.

The SAS In-Memory Statistics package includes the following software components:

SAS LASR Analytic Server
: a scalable, analytic platform that provides a secure, multi-user environment for concurrent access to in-memory data.

SAS Studio
provides an interactive web-based development application that enables you to write and submit SAS programs.

IMSTAT procedure
manages the in-memory data and SAS LASR Analytic Server instances and performs complex analytics on the in-memory data.

RECOMMEND procedure
manages tasks for a recommender system, which rates items such as movies, music, books, and so on.

SAS/ACCESS Interface to Hadoop
a SAS/ACCESS engine that enables you to interact with Hadoop through HiveServer2 or through fixed-length record (binary) data, delimited text, or XML-encoded text that is stored in HDFS. See "SAS/ACCESS Interface to Hadoop" on page 36.

Why Use SAS In-Memory Statistics?

- All mathematical calculations are performed in memory. The in-memory environment eliminates costly data movement and persists data in memory for the entire analytic session. This significantly reduces data latency and provides rapid analysis. The data is read once and held in memory for multiple processes.
- SAS In-Memory Statistics enables interactive programming access so that multiple users can analyze Hadoop data at the same time and extremely quickly.
- You can use statistical algorithms and machine-learning techniques to uncover patterns and trends in the Hadoop data.
- You can analyze unstructured and structured data using a wide range of text analysis techniques.
- You can generate personalized, meaningful recommendations in real time with a high level of customization.
- SAS In-Memory Statistics supports parallel BY-group processing.

What Is Required?

- You must license SAS In-Memory Statistics. The package includes SAS LASR Analytic Server, SAS/ACCESS Interface to Hadoop, SAS Studio, IMSTAT procedure, RECOMMEND procedure, SAS/GRAPH, SAS/STAT, and the current release of Base SAS 9.4.
- The SAS LASR Analytic Server must be co-located with a Hadoop cluster that has been configured with the services from SAS Plug-ins for Hadoop.
- The SAS LASR Analytic Server runs on a Linux x64 operating system only.
- To load data into memory, you can use the LASR procedure to load data in parallel or the SAS LASR Analytic Server engine to load data serially. Or, if your Hadoop cluster has data in Hive or HiveServer2, you can use SAS/ACCESS Interface to Hadoop to load the data into memory.
- SASHDAT and CSV files are loaded in parallel from a co-located cluster. To use SAS/ACCESS Interface to Hadoop to load data into memory in parallel and in formats other than SASHDAT and CSV, the SAS Embedded Process must be installed on the Hadoop cluster.

- SAS/ACCESS Interface to Hadoop resides on the SAS client machine that you use for submitting SAS programs. When you run a program that uses the LASR procedure and data from a SAS/ACCESS Interface to Hadoop libref, the server communicates with the remote embedded process to transfer data from Hadoop to the SAS LASR Analytic Server.
- To use SAS Studio, your user ID must be configured for passwordless SSH to the Hadoop cluster machines. Make sure that you have passwordless SSH access from the machine that hosts SAS Studio to the machines in the Hadoop cluster.

More Information

- For information about the SAS LASR Analytic Server, including the LASR procedure, IMSTAT procedure, RECOMMEND procedure, SASIOLA engine, and SASHDAT engine, see *SAS LASR Analytic Server: Reference Guide*.
- For SAS LASR Analytic Server deployment, installation, and configuration information , see *SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide*.
- For information about installing the SAS Embedded Process, see *SAS In-Database Products: Administrator's Guide*.
- For instructions about how to configure SAS/ACCESS Interface to Hadoop, see *SAS Hadoop Configuration Guide for Base SAS and SAS/ACCESS*.
- For an overview of SAS Studio and specific instructions about its use, see *SAS Studio: User's Guide*.

SAS High-Performance Analytics Products

What Are the SAS High-Performance Analytics Products?

The SAS High-Performance Analytics products enable you to execute high-performance procedures in a scalable, distributed, in-memory processing environment. The procedures include statistics, data mining, text mining, econometrics, and optimization capabilities.

The SAS High-Performance Analytics products are engineered to run in a distributed mode using a cluster of machines. When high-performance procedures execute in distributed mode, several nodes in a distributed computing environment are used for calculations. Data is distributed across the machines in a cluster, and the massive computing power of the cluster is used to solve a single large analytic task. Distributed mode enables analytical computations to be performed simultaneously on multiple machines in the cluster and across multiple, concurrently scheduled threads on each machine. In distributed computing environments, these procedures exploit parallel access to data using all of the cores and huge amounts of memory that are available.

TIP High-performance procedures can also be run in single-machine mode. Single-machine mode means multithreading is done on the client machine. The procedures use the number of cores on the client machine to determine the number of concurrent threads. To run the high-performance procedures in single-machine mode, you do not need to license the SAS High-Performance Analytics products or install and configure the SAS High-Performance Analytics environment.

SAS High-Performance Analytics products are the following:

SAS High-Performance Data Mining
: includes high-performance data mining procedures that enable you to analyze large volumes of diverse data by using a drag-and-drop interface and powerful descriptive, predictive, and machine-learning methods. A variety of modeling techniques, including random forests, support vector machines, neural networks, clustering, and so on, are combined with data preparation, data exploration, and scoring capabilities.

SAS High-Performance Econometrics
: includes high-performance econometric procedures that provide econometric modeling tools. The econometric modeling methods include the regression model for count data, a model for the severity of losses or other events, and a regression model for qualitative and limited dependent variables.

SAS High-Performance Optimization
: includes high-performance features of optimization procedures that are useful for certain classes of linear, mixed-integer linear, and nonlinear problems. Key tasks, including individual optimizations for algorithms such as multistart, decomposition, and option tuning, as well as global and local search optimization, are executed in parallel.

SAS High-Performance Statistics
: includes high-performance statistical procedures that provide predictive modeling methods. Predictive modeling methods include regression, logistic regression, generalized linear models, linear mixed models, nonlinear models, and decision trees. The procedures provide model selection, dimension reduction, and identification of important variables whenever this is appropriate for the analysis.

SAS High-Performance Text Mining
: includes high-performance text mining procedures that analyze large-scale textual data. You can gain quick insights from large unstructured data collections that involve millions of documents, emails, notes, report snippets, social media sources, and so on. Support is included for parsing, entity extraction, automatic stemming, synonym detection, topic discovery, and singular value decomposition (SVD).

Why Use the SAS High-Performance Analytics Products?

- All available computing resources are used to perform faster statistical modeling and model selection. You get finer, more accurate results to drive new opportunities for your organization.
- All data (including unstructured) is used with advanced modeling techniques.
- The high-performance analytics products can evaluate many alternative scenarios, quickly detect changes in volatile markets, and make timely, optimal recommendations.
- Analytical professionals can take full advantage of the in-memory infrastructure to solve the most complex problems without architecture constraints.
- SAS High-Performance Analytics products provide in-memory capabilities so that you can develop superior analytical models using all data, not just a sample of the data. These products load data into memory in parallel and apply complex analytical algorithms to the distributed data in memory.
- Because each process is multithreaded, the high-performance procedures maximize speed by maximizing parallel processing. Each of the multiple nodes runs a multithreaded process, and all of the data is loaded and processed in memory.

- All high-performance procedures are multithreaded and can exploit all available cores, whether on a single machine or in a distributed computing environment.
- You can execute the high-performance procedures on the SAS LASR Analytic Server in an in-memory environment. The data is loaded into memory for distributed processing and remains in memory for simultaneous access until the analytic processing completes.

What Is Required?

- You must license the current release of Base SAS 9.4.
- You must license SAS/ACCESS Interface to Hadoop.
- For SAS High-Performance Data Mining, you must license the product and SAS Enterprise Miner.
- For SAS High-Performance Text Mining, you must license the product, SAS Enterprise Miner, and SAS Text Miner.
- For SAS High-Performance Statistics, you must license the product and SAS/STAT.
- For SAS High-Performance Econometrics, you must license the product and SAS/ETS.
- For SAS High-Performance Optimization, you must license the product and SAS/OR.
- The SAS High-Performance Analytics product must be installed on the Hadoop cluster.
- The SAS High-Performance Analytics environment must be installed and configured on the Hadoop cluster.
- The SAS Embedded Process must be installed and configured on the Hadoop cluster.
- To work with SASHDAT files, Hadoop must be deployed on the same machine as each high-performance software. The Hadoop cluster must be configured with the services from SAS Plug-ins for Hadoop.

More Information

- For SAS High-Performance Data Mining, see *SAS Enterprise Miner: High-Performance Procedures*.
- For SAS High-Performance Econometrics, see *SAS/ETS User's Guide: High-Performance Procedures*.
- For SAS High-Performance Statistics, see *SAS/STAT User's Guide: High-Performance Procedures*.
- For SAS High-Performance Text Mining, see *SAS Text Miner: High-Performance Procedures*.
- For high-performance utility procedures, see *Base SAS Procedures Guide: High-Performance Procedures*.
- To install and configure the SAS High-Performance Analytics environment, see *SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide*.
- To install and configure the SAS Embedded Process, see *SAS In-Database Products: Administrator's Guide*.

- For information about the SAS LASR Analytic Server, see *SAS LASR Analytic Server: Reference Guide*.

SAS High-Performance Risk

What Is SAS High-Performance Risk?

SAS High-Performance Risk is a financial portfolio management solution that enables you to price very large portfolios at the current market state and for a large number of simulated market states. The solution can aggregate values across market states and compute risk measures on demand based on the ad hoc hierarchy that you request through a risk exploration. The solution can also be used for on-demand stress testing.

SAS High-Performance Risk includes the following software components:

- A user interface to explore project results and perform further analysis.
- HPEXPORT procedure to export SAS Risk Dimensions project specifications and static data to a format that can be used by the HPRISK procedure.
- HPRISK procedure to run the analysis projects on an analytics cluster or on a single computer system with multiple CPUs.
- An interface to SAS Event Stream Processing, which can receive data from high-speed sources, including market and reference data feeds. See “SAS Event Stream Processing” on page 49.

Why Use SAS High-Performance Risk?

SAS High-Performance Risk provides the following features:

- distributed processing on an analytics cluster
- multithreading, which increases responsiveness and concurrency
- distributed in-memory analytics to reduce the I/O burden and computational run times

What Is Required?

- You must license SAS High-Performance Risk. The package includes a web browser, the current release of Base SAS 9.4, SAS/ACCESS Interface to Hadoop, SAS Management Console, and the SAS Risk Dimensions client.
- SAS High-Performance Risk requires the SAS Intelligence Platform. The system administrator must install and configure the required SAS Intelligence Platform software. In addition, the system administrator must use SAS Management Console to maintain metadata for servers, users, and other global resources that are required by SAS High-Performance Risk.
- SAS High-Performance Risk and the SAS High-Performance Analytics environment must be installed and configured on the same Hadoop cluster. You can configure Hadoop to use SAS Plug-ins for Hadoop to enable the use of SASHDAT files.

More Information

- For information about how to use SAS High-Performance Risk, see *SAS High-Performance Risk: User's Guide*.
- For information about how to use the HPEXPORT and HPRISK procedures, see the *SAS Risk Dimensions and SAS High-Performance Risk: Procedures Guide*.
- For installation and configuration information, see *SAS High-Performance Risk: Administrator's Guide*.
- To install and configure the SAS High-Performance Analytics environment, see *SAS High-Performance Analytics Infrastructure: Installation and Configuration Guide*.

SAS In-Database Technology

What Is SAS In-Database Technology?

SAS In-Database Technology enables you to execute certain SAS processing in Hadoop. The in-database processing uses the distributed processing capabilities of MapReduce to process requests and eliminates costly data movement. In addition, in-database processing makes information more secure because the data never leaves the data source.

Why Use SAS In-Database Technology?

- You can submit the following SAS procedures to execute in Hadoop. The procedures are translated into HiveQL and processed using Hive or HiveServer2.
  - FREQ procedure
  - MEANS procedure
  - RANK procedure
  - REPORT procedure
  - SORT procedure
  - SUMMARY procedure
  - TABULATE procedure
  - TRANSPOSE procedure
- You can submit certain SAS DATA step programs to process in Hadoop. SAS determines when the code is appropriate for MapReduce. If it is appropriate, the code is executed in parallel using the data in HDFS.
- You can submit DS2 threaded programs to process in Hadoop. DS2 is a SAS proprietary programming language for table manipulation that executes in parallel in HDFS. Examples of DS2 threaded programs include large transpositions, computationally complex programs, scoring models, and BY-group processing.
- You can execute scoring models in Hadoop. The scoring models, developed by SAS Enterprise Miner or SAS Factory Miner, are translated into scoring files and stored in an HDFS directory. The scoring files are used by a MapReduce function to run the scoring models in parallel.

What Is Required?

- You must license the current release of Base SAS 9.4.
- You must license SAS/ACCESS Interface to Hadoop. All in-database processing code must include the SAS/ACCESS Interface to Hadoop LIBNAME statement to connect to the Hadoop cluster.
- A set of Hadoop JAR and configuration files must be collected and copied to the SAS client machine.
- To submit PROC TRANSPOSE, DATA step programs, DS2 threaded programs, and scoring models to execute in Hadoop, the SAS Embedded Process must be installed on the Hadoop cluster.
- To run in-database procedures, you must set SQLGENERATION= to the appropriate DBMS name using the system option or the LIBNAME option. For example, DBMS='hadoop' in the system option sets Hadoop as the data source.
- To submit PROC TRANSPOSE, you must license the SAS In-Database Code Accelerator for Hadoop.
- To execute the DATA step in Hadoop, you must set the DSACCEL= system option to ANY, use the same libref for the input and output files, and follow the DATA statement immediately by the SET statement.
- To submit DS2 threaded programs, you must license the SAS Data Loader, which includes the SAS In-Database Code Accelerator for Hadoop. To submit the DS2 program from your SAS session, use the DS2 procedure. In addition, either the PROC DS2 DS2ACCEL= option must be set to YES or the DS2ACCEL= system option must be set to ANY.
- To run scoring models, you must license SAS Enterprise Miner or SAS Factory Miner and the SAS Scoring Accelerator or you can license SAS Model Manager. Use the LIBNAME statement to specify the location of the data, metadata, and temporary data.

More Information

- For more information about using SAS In-Database Technology for submitting in-database procedures, the DATA step, DS2 threaded programs (SAS In-Database Code Accelerator), and scoring models (SAS Scoring Accelerator), see *SAS In-Database Products: User's Guide*.
- For information about the in-database deployment package for Hadoop, see the chapter "Administrator's Guide for Hadoop," in *SAS In-Database Products: Administrator's Guide*.
- For information about PROC FREQ, see *Base SAS Procedures Guide: Statistical Procedures*.
- For information about DS2, MEANS, RANK, REPORT, SORT, SUMMARY, TABULATE, and TRANSPOSE procedures, see *Base SAS Procedures Guide*.
- For more information about using the SQLGENERATION= option, see *SAS/ACCESS for Relational Databases: Reference*.

Chapter 6

Execute Models

SAS Scoring Accelerator for Hadoop

What Is SAS Scoring Accelerator for Hadoop?

SAS Scoring Accelerator for Hadoop supports executing scoring models in a Hadoop cluster. The functionality is provided as an add-on to the DS2 language.

Why Use SAS Scoring Accelerator?

The SAS Scoring Accelerator for Hadoop translates scoring models developed by SAS Enterprise Miner, SAS Factory Miner, or SAS/STAT into Hadoop functions (scoring files) that are stored in HDFS. Scoring files are then used by MapReduce to run the scoring model in the Hadoop cluster. The scoring process is performed in Hadoop, which eliminates the need to extract data. Using the parallel processing capabilities of Hadoop yields higher model-scoring performance and faster access to insights.

What Is Required?

- To use the SAS Scoring Accelerator for Hadoop, see "SAS In-Database Technology" on page 29.

Chapter 7

Manage Data

SAS Data Loader for Hadoop

What Is SAS Data Loader for Hadoop?

SAS Data Loader for Hadoop makes it easier to move, cleanse, and analyze data in Hadoop. The SAS Data Loader for Hadoop web application provides an interactive interface that guides you through the process of creating directives. You run the directives on a SAS Workspace Server, which executes generated code, sends code to Hadoop, and receives responses from Hadoop. During execution, directives provide access to generated code, log information, error messages, and results as they become available. You can save directives, update them, and execute them, as needed.

Multiple users can access the SAS Data Loader for Hadoop web application. The web application uses the SAS Web Infrastructure Platform to connect to SAS servers, to a Hadoop cluster, and to network database management servers. The SAS Metadata Server manages access to data sources and software capabilities for individuals and groups. SAS software is deployed to each node in the Hadoop cluster.

Why Use SAS Data Loader for Hadoop?

SAS Data Loader for Hadoop provides a point-and-click interface for profiling, managing, cleansing, and copying data to and from Hadoop. Power users can create directives that run SAS or HiveQL programs in Hadoop. You can use directives to perform tasks such as the following:

- Browse data
- Copy data to and from Hadoop
- Transform and transpose data
- Delete rows from source tables

- Cleanse data
- Sort data and remove duplicate rows from tables
- Query or join data
- Match-merge data
- Define rules to cluster similar records into groups
- Profile data
- Run user-written code
- Chain directives
- Manage and reuse directives
- Load data to SAS LASR Analytic Server

What Is Required?

- You must license SAS Data Loader for Hadoop.
- The following SAS software, installed on the Hadoop cluster, supports features of SAS Data Loader for Hadoop:
  - SAS Quality Knowledge Base, which supports data cleansing capabilities in Hadoop
  - SAS Embedded Process software, which runs SAS programs in Hadoop
  - SAS Data Quality Accelerator, which runs data quality capabilities in Hadoop
  - SAS Data Loader for Hadoop Spark Engine, which executes data integration and data quality tasks in Apache Spark

More Information

- For information about how to use SAS Data Loader for Hadoop, see *SAS Data Loader for Hadoop: User's Guide*.
- For information about how to install SAS Data Loader for Hadoop, see *SAS Data Loader for Hadoop: Installation and Configuration Guide*.

SAS Data Quality Accelerator for Hadoop

What Is SAS Data Quality Accelerator for Hadoop?

The SAS Data Quality Accelerator for Hadoop provides in-database data quality operations in a Hadoop cluster. The SAS Data Quality Accelerator for Hadoop optimizes access to the SAS Quality Knowledge Base in the cluster.

Why Use SAS Data Quality Accelerator?

The **Cleanse Data** directive in SAS Data Loader for Hadoop includes the following data quality transforms:

- identification analysis
- pattern analysis
- gender analysis
- parse data
- field extraction
- generate match codes
- standardize data
- change case

The SAS Data Quality Accelerator for Hadoop is used in the **Profile Data** directive.

What Is Required?

- You must license SAS Data Quality Accelerator for Hadoop.

SAS In-Database Code Accelerator for Hadoop

What Is SAS In-Database Code Accelerator for Hadoop?

SAS In-Database Code Accelerator for Hadoop supports the execution of SAS code in a Hadoop cluster. The functionality is provided as an add-on to the DS2 language.

Why Use SAS In-Database Code Accelerator?

The SAS In-Database Code Accelerator executes massively parallel DS2 programs and user-written DS2 expressions in Hadoop in coordination with Hive and HDFS. Examples of DS2 threaded programs include transpositions, match-merge, survivorship, scoring models, and BY-group processing. User-written DS2 expressions can generate new columns of target data in manage columns transformations.

What Is Required?

- You must license SAS Data Loader for Hadoop.
- You must license SAS/ACCESS Interface to Hadoop.

SAS/ACCESS Interface to Hadoop

What Is SAS/ACCESS Interface to Hadoop?

SAS/ACCESS Interface to Hadoop enables you to access Hadoop data through Hive and HiveServer2 and from HDFS. You use SAS/ACCESS Interface to Hadoop with SAS applications to access Hadoop data as SAS data sets without requiring specific Hadoop skills like writing MapReduce code.

SAS/ACCESS Interface to Hadoop works like other SAS engines. That is, you execute a LIBNAME statement to assign a libref and specify the engine. You use that libref throughout the SAS session where a libref is valid. In the LIBNAME statement, you specify the Hadoop server connection information.

Here is an example of a LIBNAME statement that connects to a Hadoop server. The LIBNAME statement assigns the libref Myhdp to the Hadoop cluster, specifies the Hadoop engine, and specifies the Hadoop server connection options.

```
libname myhdp hadoop port=100000 server=cdlserv02 user=sasabc password=hadoop;
```

Why Use SAS/ACCESS Interface to Hadoop?

- SAS/ACCESS Interface to Hadoop provides a bridge to Hadoop data so that you can run your favorite SAS user interface.
- SAS/ACCESS Interface to Hadoop supports the SQL pass-through facility, which enables SQL code to be passed to the Hadoop cluster for processing. Explicit SQL pass-through passes native HiveQL directly to the Hadoop cluster for processing. Implicit SQL pass-through translates the SQL code (generated by SAS) to HiveQL, which is then passed to the Hadoop cluster.
- SAS/ACCESS Interface to Hadoop translates Hadoop data to the appropriate SAS data type for processing with SAS.

What Is Required?

- You must license the current release of Base SAS 9.4.
- You must license SAS/ACCESS Interface to Hadoop.
- To connect to a Hadoop cluster, you must make the Hadoop cluster configuration files and Hadoop JAR files accessible to the SAS client machine. Use the SAS Deployment Manager, which is included with each SAS software order, to copy the configuration files and JAR files to the SAS client machine that connects to Hadoop. The SAS Deployment Manager automatically sets the SAS_HADOOP_CONFIG_PATH and SAS_HADOOP_JAR_PATH environment variables to the directory path.
- To connect to a Hadoop cluster using WebHDFS, you must set the SAS_HADOOP_RESTFUL environment variable to the value 1. In addition, the hdfs-site.xml Hadoop cluster configuration file must include the properties for the WebHDFS location.
- For HDFS operations, SAS/ACCESS Interface to Hadoop requires access to the Hadoop server that runs on the Hadoop cluster NameNode, which is usually on port 8020.

More Information

- For information about how to use SAS/ACCESS Interface to Hadoop, including the LIBNAME statement syntax, see *SAS/ACCESS for Relational Databases: Reference*.
- For instructions about how to configure SAS/ACCESS Interface to Hadoop, including information about configuring Hadoop JAR files and configuration files using the SAS Deployment Manager, see *SAS Hadoop Configuration Guide for Base SAS and SAS/ACCESS*.

SAS/ACCESS Interface to HAWQ

What Is SAS/ACCESS Interface to HAWQ?

SAS/ACCESS Interface to HAWQ provides direct, transparent access to the Pivotal HAWQ SQL engine from your SAS session. SAS/ACCESS Interface to HAWQ enables you to interact with HBase through the SAS LIBNAME statement and the SQL pass-through facility. You can use various LIBNAME statement options and data set options to control the data that is returned to the SAS client machine.

SAS/ACCESS Interface to HAWQ works like other SAS engines. That is, you execute a LIBNAME statement to assign a libref and specify the engine. You use that libref throughout the SAS session where a libref is valid. In the LIBNAME statement, you specify the HAWQ server connection information.

Here is an example of a LIBNAME statement that connects to a HAWQ server. The LIBNAME statement assigns the libref Myhwq to the Hadoop cluster, specifies the HAWQ engine, and specifies the HAWQ server connection options.

```
libname myhwq hawq server=hwq04 db=customers port=5432 user=hwqusr1 pw=hwqpwd1;
```

Why Use SAS/ACCESS Interface to HAWQ?

- SAS/ACCESS Interface to HAWQ supports the SQL pass-through facility, which enables SQL code to be passed to Hadoop for processing. Explicit SQL pass-through passes native HiveQL that you provide directly to the Hadoop cluster for processing. Implicit SQL pass-through translates the SQL code (generated by SAS) to HiveQL, which is then passed to the Hadoop cluster.
- SAS/ACCESS Interface to HAWQ supports bulk loading, which is much faster than inserting.
- SAS/ACCESS Interface to HAWQ is supported on the following platforms: AIX, HP-UX for Itanium, 64-bit Linux, Solaris for SPARC, 64-bit Solaris, 32-bit Microsoft Windows, and 64-bit Microsoft Windows.

What Is Required?

- You must license the current release of Base SAS 9.4.
- You must license SAS/ACCESS Interface to HAWQ.

More Information

- For information about how to use SAS/ACCESS Interface to HAWQ, including the LIBNAME statement syntax, see *SAS/ACCESS for Relational Databases: Reference*.

SAS/ACCESS Interface to Impala

What Is SAS/ACCESS Interface to Impala?

SAS/ACCESS Interface to Impala provides direct, transparent access to Impala from your SAS session. SAS/ACCESS Interface to Impala enables you to interact with HDFS through the SAS LIBNAME statement and the SQL pass-through facility. You can use various LIBNAME statement options and data set options to control the data that is returned to the SAS client machine.

SAS/ACCESS Interface to Impala works like other SAS engines. That is, you execute a LIBNAME statement to assign a libref and specify the engine. You use that libref throughout the SAS session where a libref is valid. In the LIBNAME statement, you specify the Impala server connection information.

Here is an example of a LIBNAME statement that connects to an Impala server. The LIBNAME statement assigns the libref Myimp to the Hadoop cluster, specifies the Impala engine, and specifies the Impala server connection options.

```
libname myimp impala server=sascldserv02 user=myusr1 password=mypwd1;
```

Why Use SAS/ACCESS Interface to Impala?

- You can use SAS/ACCESS Interface to Impala to read and write data to and from Hadoop as if it were any data source.
- SAS/ACCESS Interface to Impala lets you run SAS procedures against data that is accessible by Impala and returns the results to SAS.
- By interacting with Impala, which bypasses MapReduce, you gain low-latency response times and work faster.
- SAS/ACCESS Interface to Impala supports the SQL pass-through facility, which enables SQL code to be passed to Hadoop for processing. Explicit SQL pass-through passes native HiveQL directly to the Hadoop cluster for processing. Implicit SQL pass-through translates the SQL code (generated by SAS) to HiveQL, which is then passed to the Hadoop cluster.
- SAS/ACCESS Interface to Impala supports bulk loading, which is much faster than inserting.

What Is Required?

- You must license the current release of Base SAS 9.4.
- You must license SAS/ACCESS Interface to Impala.
- When bulk loading, you can connect to the Hadoop cluster through the Java API or using WebHDFS or HttpFS.
  - To connect to a Hadoop cluster using the Java API, the Hadoop JAR files must be copied to a directory that is accessible to the SAS client machine. You must set the SAS_HADOOP_JAR_PATH environment variable to the directory path for the Hadoop JAR files.

- To connect to a Hadoop cluster using WebHDFS or HttpFS, you must set the value of the SAS_HADOOP_RESTFUL environment variable to 1. In addition, the hdfs-site.xml Hadoop cluster configuration file must include the properties for the WebHDFS or HttpFS location.

  Note: When bulk loading using WebHDFS, Kerberos authentication is not honored.

- SAS/ACCESS Interface to Impala is supported on AIX, 64-bit Linux, and 64-bit Microsoft Windows.

More Information

- For information about how to use SAS/ACCESS Interface to Impala, including the LIBNAME statement syntax, see *SAS/ACCESS for Relational Databases: Reference*.
- For instructions about how to configure JAR files and for information about the SAS_HADOOP_RESTFUL environment variable, see *SAS Hadoop Configuration Guide for Base SAS and SAS/ACCESS*.

SAS/ACCESS Interface to Spark

What Is SAS/ACCESS Interface to Spark?

SAS/ACCESS Interface to Spark enables you to access data on a Spark cluster through SparkSQL. You use SAS/ACCESS Interface to Spark with SAS applications to access data as SAS data sets without requiring familiarity with DataFrame API.

SAS/ACCESS Interface to Spark works like other SAS engines. That is, you execute a LIBNAME statement to assign a libref and specify the engine. You use that libref throughout the SAS session in which a libref is valid. In the LIBNAME statement, you specify the Spark server connection information.

Here is an example of a LIBNAME statement that connects to a Hadoop server. The LIBNAME statement assigns the libref Myhdp to the Hadoop cluster, specifies the Hadoop engine, and specifies the Hadoop server connection options.

```
libname hdpSpark spark
   server="hdp31.spark.company.com" user="hive" password=had00p
   database="default" port=10016
   properties="sas.spark.jdbc.hadoop.version=2.6.3.0-235";
```

Why Use SAS/ACCESS Interface to Spark?

- SAS/ACCESS Interface to Spark supports the SQL pass-through facility, which enables SQL code to be passed to the Spark cluster for processing. Explicit SQL pass-through passes native SparkSQL directly to the Spark cluster for processing. Implicit SQL pass-through translates the SQL code (generated by SAS) to SparkSQL, which is then passed to the Spark cluster.
- SAS/ACCESS Interface to Spark translates Spark data to the appropriate SAS data type for processing with SAS.

What Is Required?

- You must license the current release of Base SAS 9.4.
- You must license SAS/ACCESS Interface to Spark. SAS/ACCESS Interface to Spark is supported on 64-bit Linux.
- To connect to a Spark cluster, you must make the Spark cluster configuration files and Spark JAR files accessible to the SAS client machine. Use SAS Deployment Manager, which is included with each SAS software order, to copy the configuration files and JAR files to the SAS client machine that connects to Hadoop. SAS Deployment Manager automatically sets the SAS_HADOOP_CONFIG_PATH and SAS_HADOOP_JAR_PATH environment variables to the directory path.

More Information

- For information about how to use SAS/ACCESS Interface to Spark, including the LIBNAME statement syntax, see *SAS/ACCESS for Relational Databases: Reference*.
- For instructions about how to configure SAS/ACCESS Interface to Spark, including information about configuring Hadoop JAR files and configuration files using SAS Deployment Manager, see *SAS Hadoop Configuration Guide for Base SAS and SAS/ACCESS*.

SQOOP Procedure

What Is the SQOOP Procedure?

The SQOOP procedure provides access to Apache Sqoop from a SAS session. Apache Sqoop transfers data between a database and HDFS.

Why Use the SQOOP Procedure?

PROC SQOOP enables you to submit Sqoop commands to your Hadoop cluster from a SAS session. The Sqoop commands are passed to the Hadoop cluster using Oozie.

What Is Required?

- You must license the current release of Base SAS 9.4.
- You must license SAS/ACCESS Interface to Hadoop.
- The Hadoop cluster must be configured to support Oozie. See your Hadoop documentation for instructions.
- To use a database with Sqoop, you must download the corresponding connectors or JDBC drivers into the Oozie Sqoop ShareLib. See your Hadoop documentation for instructions.
- You must define and set the SAS_HADOOP_CONFIG_PATH environment variable to the directory that contains the custom Hadoop cluster configuration files.

- The SAS_HADOOP_RESTFUL environment variable must be set to 1, and either WebHDFS or HttpFS must be enabled.

More Information

- For information about how to use PROC SQOOP, including syntax and instructions to set up Sqoop, see the SQOOP procedure in *Base SAS Procedures Guide*.
- For information about the SAS_HADOOP_CONFIG_PATH environment variable, see *SAS Hadoop Configuration Guide for Base SAS and SAS/ACCESS*.
- For information about the SAS_HADOOP_RESTFUL environment variable, see *SAS Hadoop Configuration Guide for Base SAS and SAS/ACCESS*.

Base SAS FILENAME Statement with the Hadoop Access Method

What Is the FILENAME Statement with the Hadoop Access Method?

The FILENAME statement with the Hadoop access method enables a SAS session to access data in HDFS. The FILENAME statement associates a fileref with an external file and the Hadoop access method.

Why Use the FILENAME Statement?

The FILENAME statement reads data from and writes data to HDFS using the SAS DATA step. Using the FILENAME statement is much like submitting the HDFS commands copyFromLocal and copyToLocal.

What Is Required?

- You must license the current release of Base SAS 9.4.
- To connect to a Hadoop cluster, the following is required:
  - In SAS 9.4M3, to connect to the Hadoop cluster, the Hadoop configuration files must be copied from the specific Hadoop cluster to a physical location that is accessible to the SAS client machine. The SAS environment variable SAS_HADOOP_CONFIG_PATH must be set to the location of the Hadoop configuration files.
  - To connect to a Hadoop cluster using the Java API, Hadoop JAR files and Hadoop cluster configuration files must be available to the SAS client machine. To make these required files available, you can use the SAS Deployment Manager, or you can manually collect these files using a Hadoop tracer script. The SAS Deployment Manager automatically sets the SAS_HADOOP_JAR_PATH environment variable and the SAS_HADOOP_CONFIG_PATH environment variable to the directory paths for the files.
  - To connect to a Hadoop cluster using WebHDFS or HttpFS, you must set the value of the SAS_HADOOP_RESTFUL environment variable to 1. In addition,

the hdfs-site.xml Hadoop cluster configuration file must include the properties for the WebHDFS or HttpFS location.

- The FILENAME statement with the Hadoop access method is not supported in the z/OS operating environment.

More Information

- For more information about using the FILENAME statement, see “FILENAME statement, Hadoop Access Method” in *SAS Global Statements: Reference*.
- For information about how to configure the FILENAME statement to connect to a Hadoop cluster, see *SAS Hadoop Configuration Guide for Base SAS and SAS/ACCESS*.

Base SAS HADOOP Procedure

What Is the HADOOP Procedure?

The HADOOP procedure enables you to interact with Hadoop data by running Apache Hadoop code. PROC HADOOP interfaces with the Hadoop JobTracker, which is the service within Hadoop that controls tasks to specific nodes in the Hadoop cluster.

Why Use the HADOOP Procedure?

PROC HADOOP enables you to submit the following:

- HDFS commands
- MapReduce programs
- Pig Latin code

What Is Required?

- You must license the current release of Base SAS 9.4.
- To connect to a Hadoop cluster, the following is required:
    - To connect to a Hadoop cluster using the Java API, Hadoop JAR files and Hadoop cluster configuration files must be available to the SAS client machine. To make these required files available, you can use the SAS Deployment Manager, or you can manually collect these files using a Hadoop tracer script. The SAS Deployment Manager automatically sets the SAS_HADOOP_JAR_PATH environment variable and SAS_HADOOP_CONFIG_PATH environment variable to the directory paths for the files.
    - To connect to a Hadoop cluster using WebHDFS or HttpFS, you must set the value of the SAS_HADOOP_RESTFUL environment variable to 1. In addition, the hdfs-site.xml Hadoop cluster configuration file must include the properties for the WebHDFS or HttpFS location.

- To connect using the Apache Oozie RESTful API to submit MapReduce programs and Pig Latin code, you must set the value of the SAS_HADOOP_RESTFUL environment variable to 1. In addition, you must set the SAS_HADOOP_CONFIG_PATH environment variable to the location where the hdfs-site.xml and core-site.xml configuration files exist. The hdfs-site.xml file must include the properties for the WebHDFS location. You need to specify Oozie properties in a configuration file and you must identify the configuration file with the PROC HADOOP statement's CFG= argument.

- To submit MapReduce programs, the hdfs-site.xml file must include the properties to run MapReduce or MapReduce 2 and YARN.
- PROC HADOOP is not supported in the z/OS operating environment.

More Information

- For information about how to use PROC HADOOP, including syntax and examples, see the HADOOP procedure in *Base SAS Procedures Guide*.
- For information about how to configure PROC HADOOP to connect to a Hadoop cluster, see *SAS Hadoop Configuration Guide for Base SAS and SAS/ACCESS*.

SAS Scalable Performance Data (SPD) Engine

What Is the SPD Engine?

The SPD Engine enables you to interact with Hadoop through HDFS. Using the SPD Engine with SAS applications, you can write data, retrieve data for analysis, perform administrative functions, and even update data as an SPD data set. The SPD Engine's computing scalability provides high-performance data delivery, accessing data sets that contain billions of observations.

The SPD Engine works like other SAS engines. That is, you execute a LIBNAME statement to assign a libref and specify the engine. You use that libref throughout the SAS session where a libref is valid. In the LIBNAME statement, you specify the pathname to a directory in a Hadoop cluster. In addition, you must include the HDFS=YES argument, which specifies to connect to the specific Hadoop cluster that is defined in Hadoop cluster configuration files.

Here is an example of a LIBNAME statement that connects to a Hadoop cluster.

```
libname myspde spde '/user/abcdef' hdfs=yes;
```

Why Use the SPD Engine?

- The SPD Engine organizes data into a streamlined file format that has advantages for a distributed file system like HDFS. Data is separate from the metadata, and the file format partitions the data.
- Most existing SAS applications can run with the SPD Engine with little modification other than to the LIBNAME statement. SAS file features such as encryption, file compression, member-level locking, indexes, SAS passwords, special missing values, user-defined formats and informats, and physical ordering of returned observations are supported.

- The SPD Engine supports parallel processing. On the SAS client machine, the SPD Engine reads and writes data stored in HDFS by running multiple threads in parallel.
- To optimize the performance of WHERE processing, you can subset data in the Hadoop cluster to take advantage of the filtering and ordering capabilities of the MapReduce framework. When you submit SAS code that includes a WHERE expression, the SPD Engine submits a Java class to the Hadoop cluster as a component in a MapReduce program. Only a subset of the data is returned to the SAS client.
- The SPD Engine supports SAS Update operations for data stored in HDFS. To update data in HDFS, the SPD Engine replaces the data set's data partition file for each observation that is updated. When an update is requested, the SPD Engine re-creates the data partition file in its entirety (including all replications), and then inserts the updated data. For a general-purpose data storage engine like the SPD Engine, the ability to perform small, infrequent updates can be beneficial.

  TIP Updating data in HDFS is intended for situations when the time it takes to complete the update outweighs the alternatives.
- The SPD Engine supports distributed locking for data stored in HDFS. For the service provider, the SPD Engine uses the Apache ZooKeeper coordination service.
- You can use the SAS High-Performance Analytics procedures on an SPD Engine data set stored in HDFS, taking advantage of the distributed processing capabilities of Hadoop. The procedures use the SAS Embedded Process to submit a MapReduce program to the Hadoop cluster.
- SPD Engine data sets can be manipulated using HiveQL. SAS provides a custom Hive SerDe so that SPD Engine data sets stored in HDFS can be accessed using Hive.

What Is Required?

- You must license the current release of Base SAS 9.4.
- To connect to a Hadoop cluster using the Java API, Hadoop JAR files and Hadoop cluster configuration files must be available to the SAS client machine. To make these required files available, you can use the SAS Deployment Manager, or you can manually collect these files using a Hadoop tracer script. The SAS Deployment Manager automatically sets the SAS_HADOOP_JAR_PATH environment variable and SAS_HADOOP_CONFIG_PATH environment variable to the directory paths for the files.
- You can connect to only one Hadoop cluster at a time per SAS session. You can submit multiple LIBNAME statements to different directories in the Hadoop cluster, but there can be only one Hadoop cluster connection per SAS session.
- To use the SAS High-Performance Analytics procedures with the SPD Engine, you must install the SAS Embedded Process on the Hadoop cluster.
- Access to data in HDFS using the SPD Engine is not supported from a SAS session in the z/OS operating environment.

More Information

- For information about how to use the SPD Engine to store data in a Hadoop cluster using HDFS, including the LIBNAME statement syntax and examples, see *SAS SPD Engine: Storing Data in the Hadoop Distributed File System*.

- For instructions about how to configure the SPD Engine to connect to a Hadoop cluster, see *SAS Hadoop Configuration Guide for Base SAS and SAS/ACCESS*.

SAS Data Integration Studio

What Is SAS Data Integration Studio?

SAS Data Integration Studio is a visual design tool for building, implementing, and managing data integration processes regardless of data sources, applications, or platforms. Through its metadata, SAS Data Integration Studio provides a single point of control for managing the following resources:

- data sources, from any platform that is accessible to SAS and from any format that is accessible to SAS
- data targets, to any platform that is accessible to SAS, and to any format that is supported by SAS
- processes that specify how data is extracted, transformed, and loaded from a source to a target
- jobs that organize a set of sources, targets, and processes (transformations)
- source code that is generated by SAS Data Integration Studio
- user-written source code

Why Use SAS Data Integration Studio?

- The Hadoop Container transformation enables you to use one transformation to perform a series of steps in one connection to a Hadoop cluster. The steps can include transfers to and from Hadoop, MapReduce processing, and Pig Latin processing.
- The Hadoop File Reader transformation reads a specified file from a Hadoop cluster.
- The Hadoop File Writer transformation writes a specified file to a Hadoop cluster.
- The Hive transformation enables you to submit your own HiveQL code in the context of a job.
- The MapReduce transformation enables you to submit your own MapReduce code in the context of a job. You must create your own MapReduce program in Java and save it to a JAR file. You then specify the JAR file in the MapReduce transformation, along with some relevant arguments.
- The Pig transformation enables you to submit your own Pig Latin code in the context of a job.
- The Transfer From Hadoop transformation transfers a specified file from a Hadoop cluster.
- The Transfer To Hadoop transformation transfers a specified file to a Hadoop cluster.
- The High-Performance Analytics transformations load and unload tables on a Hadoop cluster or a SAS LASR Analytic Server. These transformations are typically used to support a SAS Analytics solution that includes both SAS Data Integration Studio and SAS LASR Analytic Server.

- The Data Loader Directive transformation enables you to run a saved directive from SAS Data Loader for Hadoop in a SAS Data Integration Studio job.

What Is Required?

- You must license the current release of Base SAS 9.4.
- You must license an offering that includes SAS Data Integration Studio (for example, SAS Data Management Standard or Advanced).
- The Hive transformation requires "SAS/ACCESS Interface to Hadoop" on page 36 or "Base SAS HADOOP Procedure" on page 43.
- The Hadoop Container, Hadoop File Reader, Hadoop File Writer, MapReduce, Pig, Transfer From Hadoop, and Transfer to Hadoop transformations require the "Base SAS HADOOP Procedure" on page 43.
- The High-Performance Analytics transformations require a SASHDAT library, SAS LASR Analytic Server library, and login credentials that are configured for passwordless secure shell (SSH) on the machines in the analytics cluster.
- You must establish connectivity to Hadoop. This includes registering the Hadoop server and the Hadoop via Hive library on the SAS Metadata Server.

More Information

- For information about the main tasks that you can perform in SAS Data Integration Studio, including data access; data integration; metadata management; data cleansing and enrichment; extract, transform, and load (ETL); and service-oriented architecture (SOA) and message queue integration, see *SAS Data Integration Studio: User's Guide*.
- See "Establishing Connectivity to Hadoop" in the *SAS Intelligence Platform: Data Administration Guide*.
- For instructions about how to configure SAS/ACCESS Interface to Hadoop and the HADOOP procedure, see *SAS Hadoop Configuration Guide for Base SAS and SAS/ACCESS*.

Chapter 8
Additional Functionality

SAS Event Stream Processing

What Is SAS Event Stream Processing?

SAS Event Stream Processing enables programmers to build applications that can quickly process and analyze a large number of continuously flowing events. Programmers can build applications using the XML or Python. Event streams are published in applications using the C, Java, or Python Publish/Subscribe APIs; connector classes; or adapter executables.

SAS Event Stream Processing provides an HDFS adapter, which resides in dfx-esp-hdfs-adapter.jar. This JAR file bundles the Java publisher and subscriber SAS Event Stream Processing clients. The subscriber receives event streams and writes them in CSV format to HDFS. The publisher reads event streams in CSV format from HDFS and injects event blocks into a source window of SAS Event Stream Processing.

Why Use SAS Event Stream Processing?

Event stream processing is complex event processing technology that is often used for mission-critical data and decision applications. It analyzes and processes events in motion (called event streams) as they are received.

SAS Event Stream Processing allows continuous analysis of data as it is received, and enables you to incrementally update intelligence as new events occur. In addition, it is scaled for performance using distributed processing and by having the ability to filter and subset events.

Event stream processing enables the user to analyze continuously flowing data over long periods of time where low-latency incremental results are important. Event stream processing applications can analyze millions of events per second, with latencies in the milliseconds.

SAS Event Stream Processing provides the following benefits:

- The ability to pass batches of real-time information (window pulsing) for performance tuning.
- An expression language for scripting complex processing logic.
- Seamless interaction with SAS solutions and capabilities such as SAS Visual Analytics and SAS High-Performance Risk.
- Windows for filtering data, procedural and pattern matching, and aggregating.
- Flexible threading by project, which enables parallel processing when needed.

What Is Required?

- You must license SAS Event Stream Processing.
- You must have knowledge of object-oriented programming terminology and understand object-oriented programming principles.

More Information

- For more information, see the SAS Event Stream Processing Help Center..

SAS Federation Server

What Is SAS Federation Server?

The SAS Federation Server is a data server that provides scalable, threaded, multi-user, and standards-based data access technology. Using SAS Federation Server, you can process and seamlessly integrate data from multiple data sources, without moving or copying the data. SAS Federation Server provides powerful querying capabilities, as well as data source management.

Why Use SAS Federation Server?

The SAS Federation Server provides the following features.

- A centralized, web-based console that makes it easy to graphically administer, monitor, and maintain connections and data caches.
- Threaded data access technology that enhances enterprise intelligence and analytical processes.
- The ability to reference data from disparate data sources with a single query, known as data federation. It also includes its own SQL syntax, FedSQL, to provide consistent functionality, independent of the underlying data source.
- Data access control with user permissions and data source security, including row-level security and data masking.
- A driver for SASHDAT, which is a Write-only driver designed for use with Hadoop on SAS LASR Analytic Server. SAS LASR Analytic Server integrates with Hadoop by storing data in HDFS. Using the SASHDAT driver, you can access the SAS LASR Analytic Server and transfer data to HDFS. Because the data volumes in HDFS are usually very large, the SASHDAT driver is not designed to read data from HDFS and transfer it back to the client.

What Is Required?

- You must license SAS Federation Server.
- You must license SAS Federation Server Driver for Apache Hive.
- To use SAS Federation Server to write SASHDAT files to HDFS, you must license SAS LASR Analytic Server. Also, SAS Federation Server Driver for SASHDAT requires SAS/SECURE to operate.

More Information

- For information about how to administer SAS Federation Server, see *SAS Federation Server: Administrator's Guide*.
- For information about how to set up data views and caches with security, see *SAS Federation Server Manager: User's Guide*.
- For information about the SAS LASR Analytic Server, see *SAS LASR Analytic Server: Reference Guide*.
- For information about FedSQL, see *SAS FedSQL Language Reference*.

SAS Grid Manager for Hadoop

What Is SAS Grid Manager for Hadoop?

SAS Grid Manager for Hadoop provides workload management, accelerated throughput, and the ability to schedule SAS analytics on your Hadoop cluster. SAS Grid Manager for Hadoop leverages YARN to manage resources and distribute SAS analytics to a Hadoop cluster running multiple applications. Oozie provides the scheduling capability for SAS workflows.

Why Use SAS Grid Manager for Hadoop?

- If you have a shared Hadoop cluster that is running multiple workloads and leveraging YARN for resource management, and you want to also run SAS analytics on this shared Hadoop cluster, SAS Grid Manager for Hadoop is required.
- Because SAS Grid Manager for Hadoop is integrated with YARN just like other SAS High-Performance technologies such as SAS High-Performance Analytics and SAS Visual Analytics, it is useful for these SAS technologies to run co-located on compute nodes next to your Hadoop data nodes and to leverage YARN to share the resources between these SAS technologies.

What Is Required?

- You must license the current release of Base SAS 9.4.
- You must license SAS/CONNECT.
- You must license SAS Grid Manager for Hadoop.
- You must use Kerberos to secure the Hadoop cluster.

More Information

For information about using a SAS grid, see *Grid Computing in SAS*.

SAS Scalable Performance Data (SPD) Server

What Is the SPD Server?

SPD Server provides a multi-user, high-performance data delivery environment that enables you to interact with Hadoop through HDFS. Using SPD Server with SAS applications, you can read and write tables and perform intensive processing (queries and sorts) on a Hadoop cluster.

Why Use the SPD Server?

- SPD Server organizes data into a streamlined file format that has advantages for a distributed file system like HDFS. Data is separate from the metadata, and the file format partitions the data.
- SPD Server supports parallel processing. On the SAS client machine, SPD Server reads and writes data stored in HDFS by running multiple threads in parallel.
- SPD Server provides a multi-user environment.
- SPD Server is a full 64-bit server that supports up to two billion columns and (for all practical purposes) unlimited rows of data. SPD Server tables are stored on disk in a format that enhances access and supports large table requirements for SAS 9.4. SPD Server cluster tables are a unique design feature that enhances managing large tables by enabling the user to create a virtual table that consists of several SPD Server tables.
- SPD Server uses access control lists (ACLs) and SPD Server user IDs to secure domain resources.
- If the Hadoop cluster supports Kerberos, SPD Server honors Kerberos ticket cache-based logon authentication and authorization as long as the Hadoop cluster configuration files are accessible.
- SPD Server supports distributed locking for data stored in HDFS. For the service provider, SPD Server uses the Apache ZooKeeper coordination service.

What Is Required?

- You must license the SAS Scalable Performance Data Server.
- To use the SPD Server with a Hadoop cluster, SPD Server must be on a Linux x64 or Microsoft Windows x64 operating system.
- To read and write to a Hadoop cluster, the SPD Server administrator must enable the SPD Server for Hadoop.

More Information

- To operate the SPD Server, see *SAS Scalable Performance Data Server: User's Guide*.
- To configure and administer the SPD Server, see *SAS Scalable Performance Data Server: Administrator's Guide*.
- To enable SPD Server to read and write to a Hadoop cluster, see *SAS Scalable Performance Data Server: Processing Data in Hadoop*.

SAS Visual Scenario Designer

What Is SAS Visual Scenario Designer?

SAS Visual Scenario Designer is a visual tool that uses data to identify events or patterns that might be associated with fraud or non-compliance. This solution enables you to gather and analyze customized data collections to create data-driven scenarios that accurately detect customer patterns. This application uses SAS LASR Analytic Server to aggregate and simulate those patterns on the input data set.

SAS Visual Scenario Designer is a middle-tier solution that is supported by SAS LASR Analytic Server. As such, SAS Visual Scenario Designer supports other client applications that you can use to create a complete investigation and detection solution.

Why Use SAS Visual Scenario Designer?

Use SAS Visual Scenario Designer to enhance the analytic power of your data, explore new data sources, investigate them, and create visualizations to uncover relevant patterns. These patterns might represent events or patterns of interest that require further investigation or reporting.

SAS Visual Scenario Designer uses a robust window-building capability to provide you with diverse and interactive detection tools. Windows can be used to feed other windows and tables to expand your exploration options. After exploring a scenario, you can activate a deployment that is based on the results. The deployment component enables you to easily change parameter values for any window or scenario in the deployment. This means that SAS Visual Scenario Designer provides near real-time exploration capability.

Visualizations can be shared easily via reports. Traditional reporting is prescriptive. That is, you know what you are looking at and what you need to convey. However, SAS Visual Scenario Designer data discovery invites you to plumb the data, its characteristics, and its relationships. This provides you with a powerful and versatile analytic tool in the SAS Fraud and Compliance solutions family.

Note: Starting in SAS 9.4M7, SAS Visual Scenario Designer is no longer available. Instead, customers should use the SAS Detection and Investigation products to modernize their implementation.

What Is Required?

- You must license SAS Visual Scenario Designer, which includes the current release of Base SAS 9.4 and SAS LASR Analytic Server.

More Information

- For more information about how to use SAS Visual Scenario Designer, see *SAS Visual Scenario Designer: User's Guide*.
- For administration of SAS Visual Scenario Designer, see *SAS Visual Scenario Designer: Administrator's Guide*.

Recommended Reading

Here is the recommended reading list for this title:

- See each SAS technology summary in this document for references to the full product documentation.
- SAS offers instructor-led training and self-paced e-learning courses. The course *Introduction to SAS and Hadoop* teaches you how to use SAS programming methods to read, write, and manipulate Hadoop data. The course *DS2 Programming Essentials with Hadoop* focuses on DS2, which is a fourth-generation SAS proprietary language for advanced data manipulation. For more information about the courses available, see SAS Training.
- *SAS and Hadoop Technology: Deployment Scenarios*

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

Apache Hadoop (Hadoop)

an open-source framework that enables the distributed processing of large data sets, across clusters of computers, using a simple programming model.

Apache Hive (Hive)

a declarative SQL-like language that presents data in the form of tables for Hadoop. Hive incorporates HiveQL (Hive Query Language) for declaring source tables, target tables, joins, and other functions to SQL that are applied to a file or set of files available in HDFS.

Apache Sqoop (Sqoop)

a command-line interface application that transfers data between Hadoop and relational databases.

Base SAS

the core product that is part of SAS Foundation and is installed with every deployment of SAS software. Base SAS provides an information delivery system for accessing, managing, analyzing, and presenting data.

big data

information (both structured and unstructured) of a size, complexity, variability, and velocity that challenges or exceeds the capacity of an organization to handle, store, and analyze it.

Cloudera Impala (Impala)

an open-source SQL query engine that provides massively parallel processing for data stored in a computer cluster on Apache Hadoop.

cluster

See computer cluster.

commodity cluster computing (commodity computing)

the use of large numbers of inexpensive computers for parallel computing to get the greatest amount of useful computation at low cost. Commodity computing involves low-performance computers working in parallel, in contrast to the use of fewer but more expensive high-performance machines. *See also* commodity hardware.

commodity computing

See commodity cluster computing.

commodity hardware

general-purpose computers that can be readily obtained from multiple vendors and that frequently incorporate components based on open standards.

computer cluster (cluster)

a set of two or more connected computers in a centralized, cohesive system that shares tasks, such as data storage and analytical computations, across the system for fast, reliable processing. A cluster can be established to achieve higher levels of performance and load distribution, or to increase reliability through redundancy.

distributed data

data that is divided and stored across multiple connected computers.

Embedded Process

See SAS Embedded Process.

Hadoop

See Apache Hadoop.

Hadoop configuration file

a file that defines how a system connects to the Hadoop cluster and provides system information.

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.

Hadoop distribution

a collection of Hadoop components such as HDFS, Hive, and MapReduce. A commercial Hadoop distribution is provided by a vendor such as Cloudera and Hortonworks.

Hadoop YARN (YARN)

a Hadoop module that serves as a resource management framework for scheduling and handling computing resources for distributed applications.

HBase

an open source, non-relational, distributed database that runs on top of HDFS, providing a fault-tolerant way of storing large quantities of sparse data.

HDFS

See Hadoop Distributed File System.

high-performance

a quality of computing performance that is characterized by significantly reduced processing time and greater throughput than that obtained by conventional means (such as sequential algorithms, single processors, and traditional databases).

Hive

See Apache Hive.

Impala

See Cloudera Impala.

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, and so on) into one file to distribute application software or libraries on the Java platform.

Java Archive

See JAR.

MapReduce

a component of Apache Hadoop, a parallel programming model for distributed processing of large data sets. The Map phase performs operations such as filtering, transforming, and sorting. The Reduce phase aggregates the output.

massively parallel processing (MPP)

the use of a large number of processors (or separate computers) to perform a set of coordinated computations in parallel.

MPP

See massively parallel processing.

node server

a computer that acts as a server in a network that uses multiple servers.

parallel execution

See parallel processing.

parallel processing (parallel execution)

a method of processing that divides a large job into multiple smaller jobs that can be executed simultaneously on multiple CPUs.

Pig

a high-level procedural language that helps manipulate data stored in HDFS. It provides a way to do ETL and basic analysis without having to write MapReduce programs.

rack server

a collection of servers that are stacked in order to minimize floor space and to simplify cabling among network components. A rack server configuration typically has a special cooling system to prevent excessive heat buildup that would otherwise occur when many power-dissipating components are confined in a small space.

SAS accelerator

a software component that supports executing SAS code in a data source.

SAS Embedded Process (Embedded Process)

a portable, lightweight execution container for SAS code that makes SAS portable and deployable on a variety of platforms.

SAS High-Performance Analytics Environment (SAS HPA Grid)

the distributed computing environment for SAS High-Performance Analytics.

SAS HPA Grid

See SAS High-Performance Analytics Environment.

SAS LASR Analytic Server

a Read-only, stateless, in-memory analytic platform that provides secure, multi-user, concurrent access to data loaded into memory in a distributed computing environment.

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.

serde

an interface that enables serialization or deserialization of one or more file formats.

Sqoop

See Apache Sqoop.

WebHDFS

an HTTP REST API that supports the complete file system interface for HDFS.

YARN

See Hadoop YARN.

Index