



THE  
POWER  
TO KNOW.

# **SAS<sup>®</sup> Data Loader 2.2 for Hadoop**

User's Guide

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2015. *SAS® Data Loader 2.2 for Hadoop: User's Guide*. Cary, NC: SAS Institute Inc.

### **SAS® Data Loader 2.2 for Hadoop: User's Guide**

Copyright © 2015, SAS Institute Inc., Cary, NC, USA

All rights reserved. Produced in the United States of America.

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

**For a web download or e-book:** Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication. The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

**NOTICE:** This documentation contains information that is proprietary and confidential to SAS Institute Inc. It is provided to you on the condition that you agree not to reveal its contents to any person or entity except employees of your organization or SAS employees. This obligation of confidentiality shall apply until such time as the company makes the documentation available to the general public, if ever.

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

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

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513–2414.

Printing 1, March 2015

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

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

With respect to CENTOS third party technology included with the vApp ("CENTOS"), CENTOS is open source software that is used with the Software and is not owned by SAS. Use, copying, distribution and modification of CENTOS is governed by the CENTOS EULA and the GNU General Public License (GPL) version 2.0. The CENTOS EULA can be found at [http://mirror.centos.org/centos/6/os/x86\\_64/EULA](http://mirror.centos.org/centos/6/os/x86_64/EULA). A copy of the GPL license can be found at <http://www.opensource.org/licenses/gpl-2.0> or can be obtained by writing to the Free Software Foundation, Inc., 59 Temple Place, Suite 330, Boston, MA 02110-1301 USA. The source code for CENTOS is available at <http://vault.centos.org/>.

With respect to open-vm-tools third party technology included in the vApp ("VMTTOOLS"), VMTTOOLS is open source software that is used with the Software and is not owned by SAS. Use, copying, distribution and modification of VMTTOOLS is governed by the GNU General Public License (GPL) version 2.0. A copy of the GPL license can be found at <http://www.opensource.org/licenses/gpl-2.0> or can be obtained by writing to the Free Software Foundation, Inc., 59 Temple Place, Suite 330, Boston, MA 02110-1301 USA. The source code for VMTTOOLS is available at <http://sourceforge.net/projects/open-vm-tools/>.

With respect to VIRTUALBOX third party technology included in the vApp ("VIRTUALBOX"), VIRTUALBOX is open source software that is used with the Software and is not owned by SAS. Use, copying, distribution and modification of VIRTUALBOX is governed by the GNU General Public License (GPL) version 2.0. A copy of the GPL license can be found at <http://www.opensource.org/licenses/gpl-2.0> or can be obtained by writing to the Free Software Foundation, Inc., 59 Temple Place, Suite 330, Boston, MA 02110-1301 USA. The source code for VIRTUALBOX is available at <http://www.virtualbox.org/>.

---

## Contents

<b>Chapter 1 • Introducing SAS Data Loader for Hadoop</b>	<b>1</b>
Hadoop for Everyone	1
How It Works	2
Implementation	3
<b>Chapter 2 • Get Started</b>	<b>5</b>
Initial Steps	5
What You Will See	6
Start, Shut Down, and Restart the vApp for SAS Data Loader	6
Open and Close the SAS Data Loader: Information Center	9
Open, Close, and Reopen SAS Data Loader for Hadoop	9
About the Execution of Jobs	10
Using the Directive Interface	10
Create and Execute a Job Using SAS Sample Data	12
Get Comfortable	14
<b>Chapter 3 • Manage Data in Hadoop</b>	<b>19</b>
Overview of Data Management Directives	19
Cleanse Data in Hadoop	20
Run a SAS Program	31
Query or Join Data in Hadoop	32
Sort and De-Duplicate Data in Hadoop	38
Transform Data in Hadoop	42
Transpose Data in Hadoop	51
<b>Chapter 4 • Profile Data in Hadoop</b>	<b>53</b>
Overview of Profile Directives	53
Profile Data	55
Saved Profile Reports	62
<b>Chapter 5 • Copy Data To and From Hadoop</b>	<b>69</b>
Overview of the Copy Data Directives	69
Copy Data to Hadoop	70
Copy Data from Hadoop	84
Load Data to LASR	95
<b>Chapter 6 • Manage Jobs</b>	<b>101</b>
Overview of Job Management Directives	101
Run Status	102
Saved Directives	104
<b>Chapter 7 • Client Administration</b>	<b>107</b>
Introduction	107
Update the vApp for SAS Data Loader	107
Troubleshoot the vApp Start Process	108
Update Kerberos Security on the vApp	109
Protect the vApp Directory	109
Troubleshoot Jobs	109
About Session Time-out	111
About Hadoop Client JAR Files and Client Configuration Files	111

Change the Version of Hadoop .....	112
Change the Hadoop Server Connection .....	112
Change the File Format of Hadoop Target Tables .....	113
Enable Logging inside the vApp .....	115
Manage Your License .....	115
Download Emergency SID Files .....	116
<b><i>Recommended Reading</i></b> .....	<b>119</b>
<b><i>Index</i></b> .....	<b>121</b>

## 1

# Introducing SAS Data Loader for Hadoop

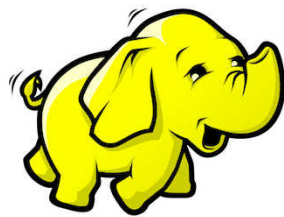
<i>Hadoop for Everyone</i> .....	1
<i>How It Works</i> .....	2
<i>Implementation</i> .....	3

---

## Hadoop for Everyone

The SAS Data Loader software opens the vast resources of Hadoop to a wider community, and adds the power of SAS to maximize the extraction of knowledge. Today, instead of requiring consultation, business analysts use this approachable wizard-based web application to perform a full range of data management tasks, all of which run directly in Hadoop.

You do not need to be an expert in Hadoop to be able to run with the elephant! You too can copy data to and from Hadoop. You too can profile, cleanse, query, transform, and analyze data in Hadoop.



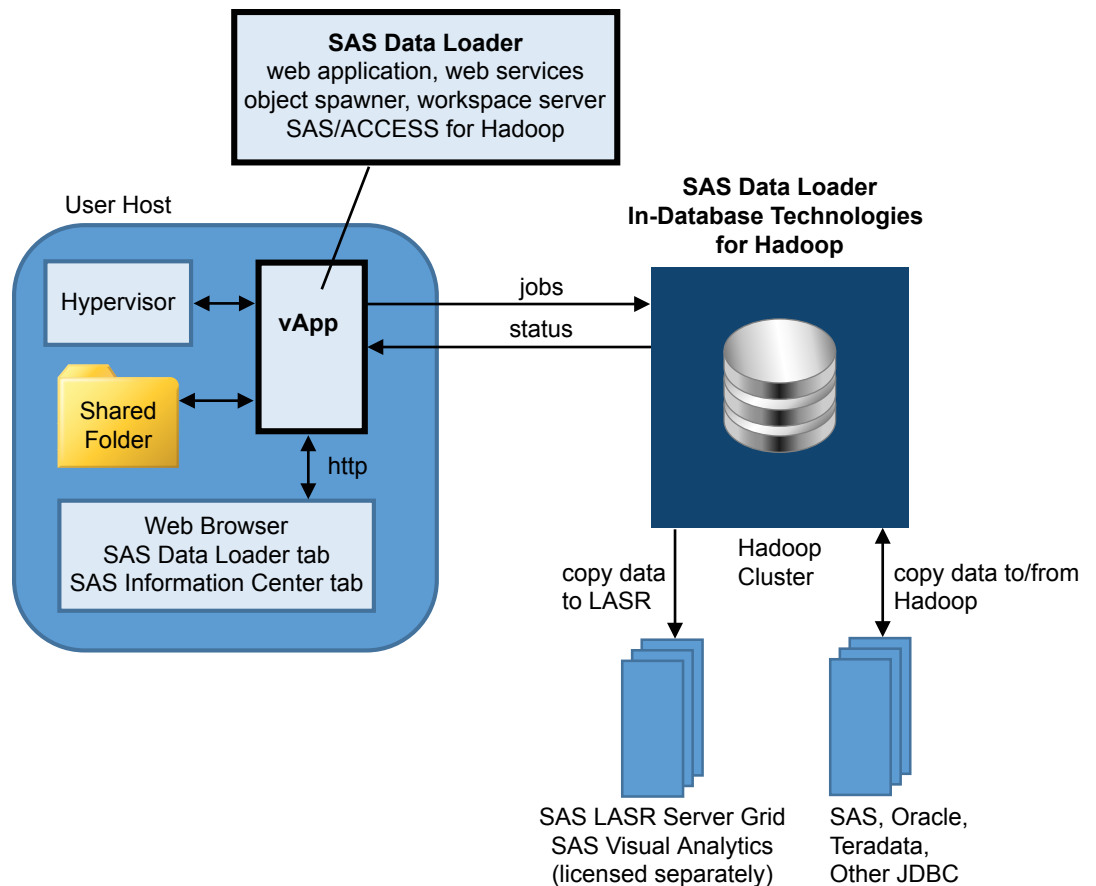
Hadoop experts can also appreciate ease-of-use. SAS Data Loader for Hadoop builds directives as jobs. Each job generates and displays executable code, which can be edited and saved for reuse. SAS DS2 programs, DS2 expressions, and HiveQL expressions can be dropped into directives to repeat execution and simplify job management.

Everyone can appreciate client software that is easy to install, configure, secure, and update. The web application for SAS Data Loader for Hadoop is delivered and runs in a virtual machine called a vApp. The vApp installs quickly, runs in isolation using a guest operating system, and updates with a single click. The vApp for SAS Data Loader is rapidly configured to use the Kerberos security system that is implemented in many Hadoop environments.

To bring the power of SAS into Hadoop, the SAS In-Database Technologies for Hadoop are deployed across the nodes of your Hadoop cluster. The in-database software enables data cleansing and embedded-process efficiency.

## How It Works

The following diagram illustrates the installed configuration of SAS Data Loader.



The SAS Data Loader for Hadoop web application runs inside the vApp. The vApp is started and managed by a hypervisor application called VMware Player Pro.

The hypervisor provides a web (HTTP) address that you enter into a web browser. The web address opens the SAS Data Loader: Information Center. The SAS Data Loader: Information Center does the following:

- starts the SAS Data Loader web application in a new browser tab.
- provides a single Settings window to configure the vApp connection to Hadoop.
- checks for available vApp software updates and installs vApp software updates.

All of the files that are accessed by the vApp reside in the Shared Folder. The Shared Folder is the only location on the user host that is accessed by the vApp. The Shared Folder contains your saved jobs, the JDBC drivers needed to

connect to external databases, and the Hadoop JAR files that were copied to the client from the Hadoop cluster.

When you create a job using a directive, the web application generates code that is then sent to the Hadoop cluster for execution. When the job is complete, the Hadoop cluster writes data to the target file and delivers log and status information to the vApp.

The SAS In-Database Technologies for Hadoop software is deployed to each node in the Hadoop cluster. The in-database technologies consist of a SAS Quality Knowledge Base for reference to data cleansing definitions, SAS Embedded Process software for code acceleration, and SAS Data Quality Accelerator software for SAS DS2 methods that pertain to data cleansing.

The SAS Data Loader for Hadoop provides the following categories of directives:

#### Copy data to and from Hadoop

Copy data as needed to and from SAS and databases outside of Hadoop. Also copy data out to SAS LASR Analytic Servers for analysis with SAS Visual Analytics and SAS Visual Statistics.

#### Manage data in Hadoop

Directives support combinations of queries, summarizations, joins, transformations, sorts, filters, column management, and de-duplication. Data quality transformations include standardization, parsing, match code generation, and identification analysis, combined with available filtering and column management to reduce the size of target tables.

#### Profile data

Profile jobs examine the quality and content of tables and produce reports. The reports are stored and managed for future reference. When you select source and target tables for your jobs, you can open the profile reports of the tables that have been profiled.

#### Manage jobs

When you create jobs, you can save them for later execution and edit. When you run jobs, the Run Status directive shows you the run status and enables you to stop execution and return the job for edit in a directive.

---

## Implementation

The implementation of SAS Data Loader for Hadoop follows these steps:

- 1 Install and configure the SAS In-Database Technologies for Hadoop.
  - Administrators (Hadoop, SAS, or System) deploy software on the nodes of a Hadoop cluster.
  - The in-database technologies are documented entirely and exclusively in the *SAS Data Loader for Hadoop: Administrator's Guide*.
- 2 Deliver site-specific files and settings to vApp installers.
  - Based on site specifics, administrators deliver Hadoop JAR files, JDBC drivers, and settings for Kerberos security, database connections, and SAS LASR Analytic Server connections.

- Administrators refer to the *SAS Data Loader for Hadoop: Administrator's Guide*. Persons installing the vApp web application refer exclusively to the *SAS Data Loader for Hadoop: vApp Deployment Guide*.

**3** Install and configure the vApp.

- Persons installing the vApp install a hypervisor, install the vApp for SAS Data Loader, and configure both.
- Persons installing the vApp refer exclusively to the *SAS Data Loader for Hadoop: vApp Deployment Guide*. vApp installation is straightforward. The configuration of the vApp takes place entirely in the user interface of the web application .

**4** Use and administer SAS Data Loader for Hadoop.

- vApp users create jobs, administer the vApp, and update the vApp using this document, the *SAS Data Loader for Hadoop: User's Guide*.
- Administrators manage the in-database SAS software using the *SAS Data Loader for Hadoop: Administrator's Guide*.



## 2

## Get Started

<b>Initial Steps</b> .....	<b>5</b>
<b>What You Will See</b> .....	<b>6</b>
<b>Start, Shut Down, and Restart the vApp for SAS Data Loader</b> .....	<b>6</b>
Overview .....	6
Use the vApp .....	7
Usage Notes for VMware Player Pro .....	8
<b>Open and Close the SAS Data Loader: Information Center</b> .....	<b>9</b>
<b>Open, Close, and Reopen SAS Data Loader for Hadoop</b> .....	<b>9</b>
<b>About the Execution of Jobs</b> .....	<b>10</b>
<b>Using the Directive Interface</b> .....	<b>10</b>
<b>Create and Execute a Job Using SAS Sample Data</b> .....	<b>12</b>
<b>Get Comfortable</b> .....	<b>14</b>
Overview .....	14
About the SAS Data Loader Directives Page .....	14
About Data Sources, Source Tables, and Target Tables .....	14
About the SAS Table Viewer .....	15
About the Sample Data Viewer .....	17
About the Code Editor .....	18

## Initial Steps

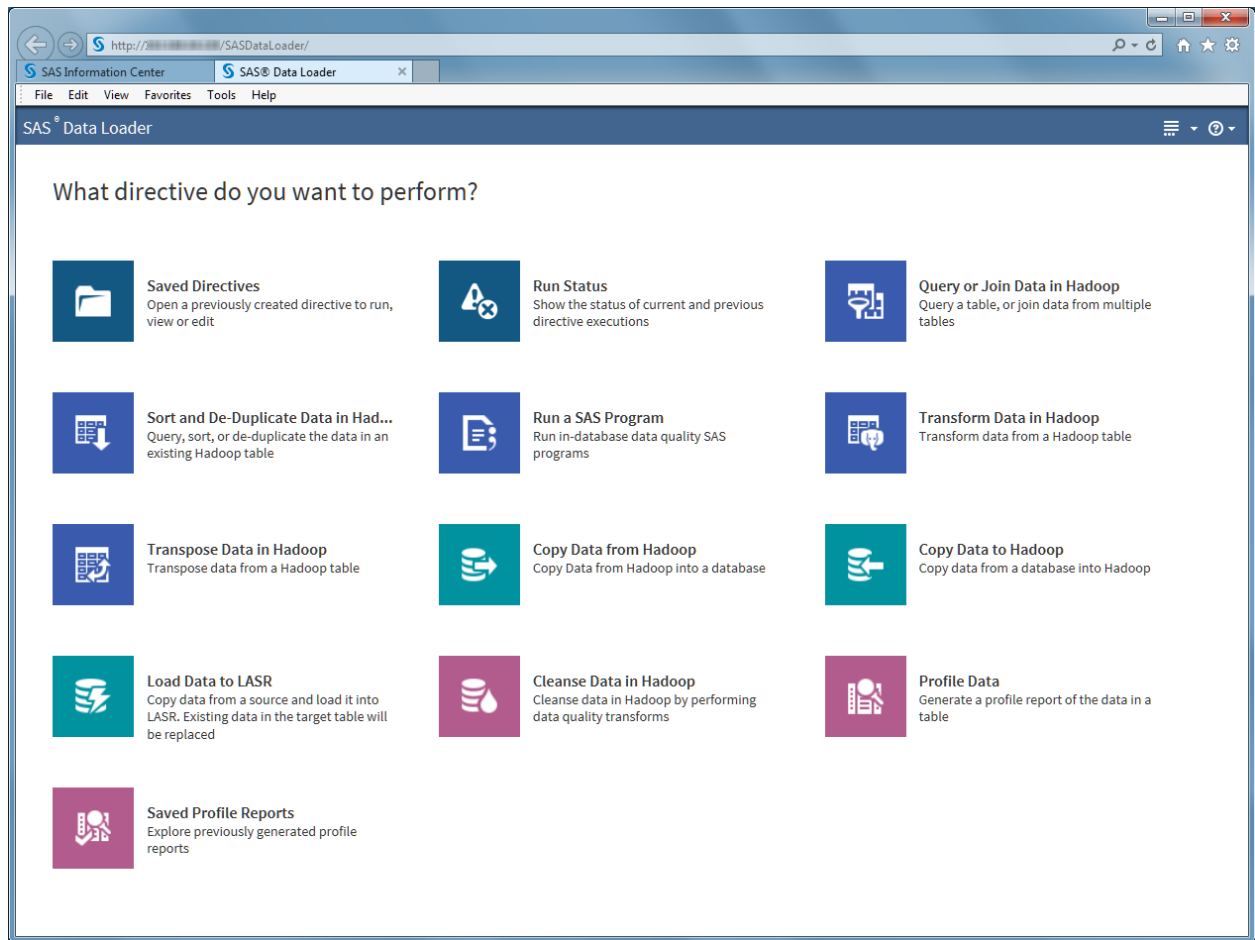
Use this chapter to learn how to start and run the SAS Data Loader for Hadoop. This chapter also familiarizes you with window layout and the basic operating characteristics of the web application.

This document assumes that your vApp client has been installed and configured using the *SAS Data Loader for Hadoop: vApp Deployment Guide*. Please complete or verify the completion of the configuration process before you begin using SAS Data Loader for Hadoop.

In addition to configuring the vApp client, an administrator needs to deploy the SAS In-Database Technologies for Hadoop across the nodes of your Hadoop cluster. Contact your Hadoop administrator as needed to determine the operational status of your Hadoop cluster. All of the installation and administration information for the in-database technologies is provided in the *SAS Data Loader for Hadoop: Administrator's Guide*.

## What You Will See

The next three topics show you how to start the vApp, open the SAS Data Loader: Information Center, and open the SAS Data Loader for Hadoop web application. The initial page of the web application provides direct access to all of the primary functions of the software.



The primary functions of SAS Data Loader for Hadoop are called directives. Directives guide you through the process of creating and running jobs in Hadoop. Click an icon to begin your journey!

## Start, Shut Down, and Restart the vApp for SAS Data Loader

### Overview

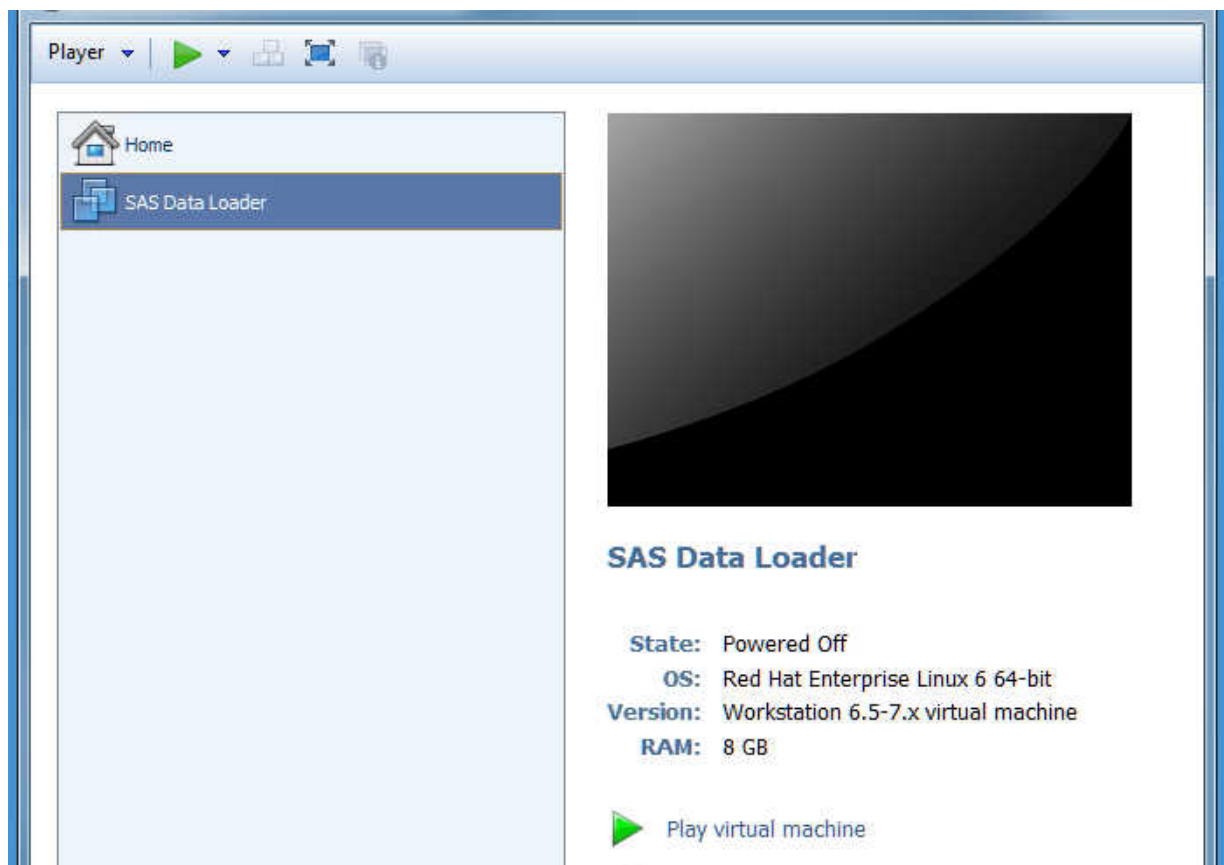
To open SAS Data Loader for Hadoop web application, you first start the vApp for SAS Data Loader and open the SAS Data Loader: Information Center. The

vApp is a virtual machine that runs a guest operating system on your client host. SAS Data Loader for Hadoop runs inside the vApp. The vApp is operated by a hypervisor, which reserves memory for the vApp and starts essential services. SAS Data Loader for Hadoop uses a hypervisor called VMware Player Pro. The hypervisor was installed and configured on your client host as specified in the *SAS Data Loader for Hadoop: vApp Deployment Guide*.

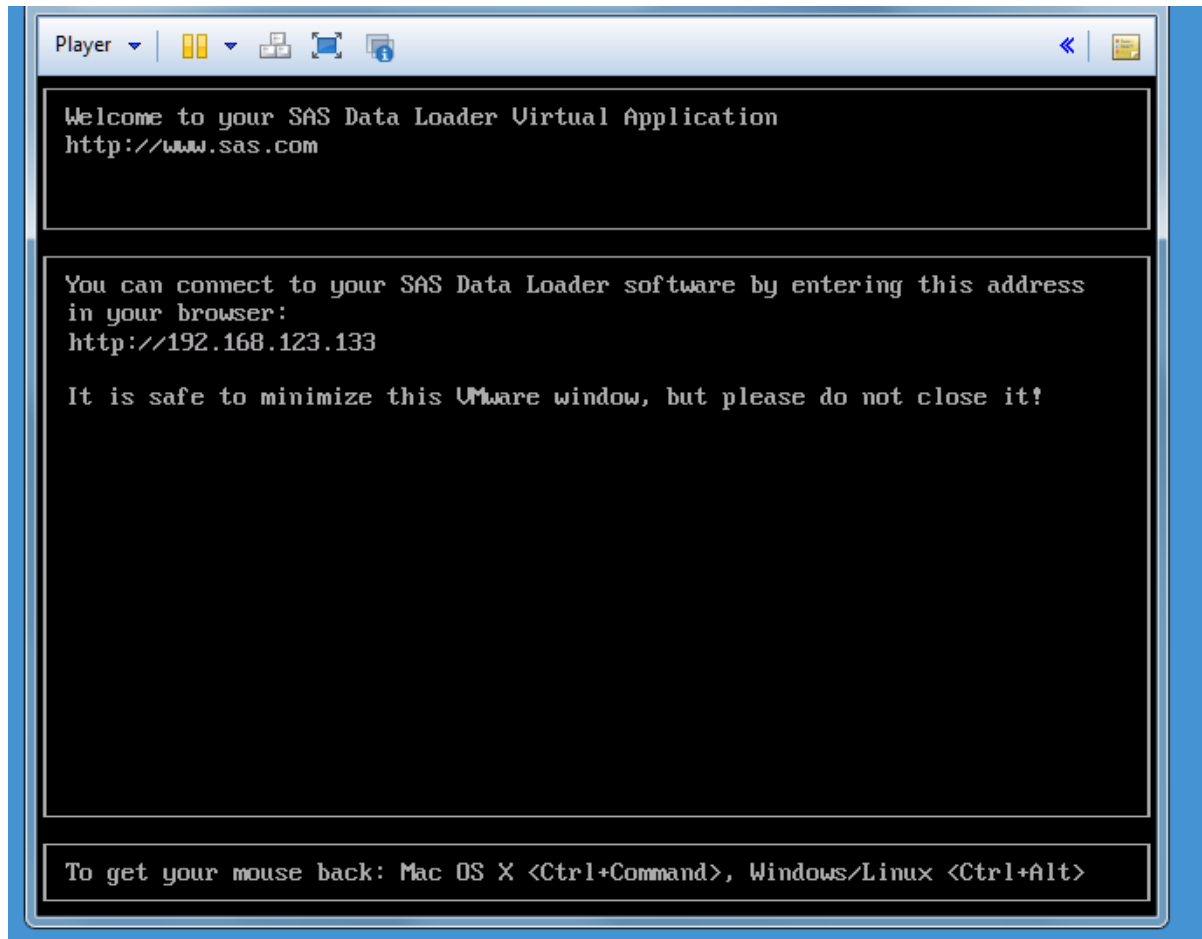
## Use the vApp

Follow these steps to start, shut down, and restart the vApp using VMware Player Pro:

- 1 Open VMware Player Pro.
- 2 Click **SAS Data Loader** and then, when it appears, click **Play virtual machine**.



- 3 VMware Player Pro requires a minute or two to play the vApp. When the vApp is ready, the VMware Player displays the message *Welcome to your SAS Data Loader Virtual Application*. You might receive an informational *Removable Devices* window. Review the information about removable devices and click **OK** to close that window.



- 4 When you power off or shut down the vApp, first [close SAS Data Loader](#) if it is open. Then, in the SAS Data Loader – VMware Player Pro window, click **Player ► Power ► Shut Down Guest**. (The term guest refers to the guest operating system that runs the vApp.)
- 5 In the VMware Player window, click **Yes** to confirm that you want to power-off the vApp.

## Usage Notes for VMware Player Pro

You can close and reopen the SAS Data Loader web application without shutting down the vApp. The vApp continues to play until you shut it down in VMware Player Pro.

Do not close the VMware Player Pro window while the vApp is playing.

Note that if the vApp is playing, and if you click in the window SAS Data Loader – VMware Player Pro, the cursor disappears. This behavior is expected; it ensures that you have to physically enter the web address in a web browser to open the SAS Data Loader: Information Center. To restore your cursor, press Ctrl+Alt.

**CAUTION! Do not pause or suspend the vApp for SAS Data Loader.** VMware Player Pro provides a capability to suspend vApps. Suspending the vApp for SAS Data Loader is not supported. In VMware Player Pro, do not select **Suspend Guest** or **Player ► Power ► Suspend Guest**. Suspending the vApp can interrupt

communications between the SAS Data Loader web client and the Hadoop cluster. To resolve a suspended vApp, select **Player ► Power ► Restart Guest**.

You need to restart the vApp after you change the version of Hadoop or configure a connection to a new database (see [Chapter 5, “Copy Data To and From Hadoop,” on page 69](#).) To restart the vApp, open VMware Player Pro and select **Player ► Power ► Restart Guest**. (The term Guest refers to the guest operating system that runs in the vApp.)

---

## Open and Close the SAS Data Loader: Information Center


The SAS Data Loader: Information Center is the launching point for the SAS Data Loader for Hadoop web application. The information center also provides important notifications, configuration settings, and access to automated updates of the vApp.

To open the information center, you first need to [start the vApp](#).

With the vApp running, you can open the information center in the following ways:

- Click the browser tab for the SAS Information Center, if it is still available.
- Double-click the desktop icon for the SAS Information Center, if it is available in Windows.
- Open the VMware Player Pro window and enter the displayed URL in a browser tab.

Features in the information center include:

- **Start SAS Data Loader** button. One click starts the web application.
- The **Notifications** section enables you to check for available vApp updates and initiate vApp updates.
- The **Resources** section provides helpful links.
- The Settings icon  opens the Settings window.
- The displayed version of Hadoop (**Hortonworks 2.1**, in this case) is specified in the Settings window.

To close the SAS Data Loader: Information Center, simply close the browser tab. You can reopen the SAS Data Loader: Information Center at any time, as described above.

---

## Open, Close, and Reopen SAS Data Loader for Hadoop

Follow these steps to open and close SAS Data Loader for Hadoop:

- 1 Open the [VMware Player Pro](#) and play the vApp for SAS Data Loader.

- 2 Open the [SAS Data Loader: Information Center](#) and then click **SAS Data Loader**. The SAS Data Loader web application opens in a new tab in your web browser.

**Note:** When starting SAS Data Loader for Hadoop, if an error occurs stating that VT-x or AMD-v is not available, see [“Troubleshoot the vApp Start Process” on page 108](#).




- 3 You can close the SAS Data Loader web browser at any time. You can open it again by clicking in the SAS Data Loader: Information Center. With the browser tab closed, running jobs continue to run and job statuses continue to be collected.
- 4 To close SAS Data Loader for Hadoop entirely, you need to [stop the vApp](#). Make sure that all jobs have been stopped before you stop the vApp, as displayed in the [“Run Status”](#) directive.

---

## About the Execution of Jobs

In SAS Data Loader for Hadoop, a job is a program that is executed on a specified Hadoop cluster. A job consists of code that accesses specified data sources at specified times.

The execution of a job in a directive follows these steps:

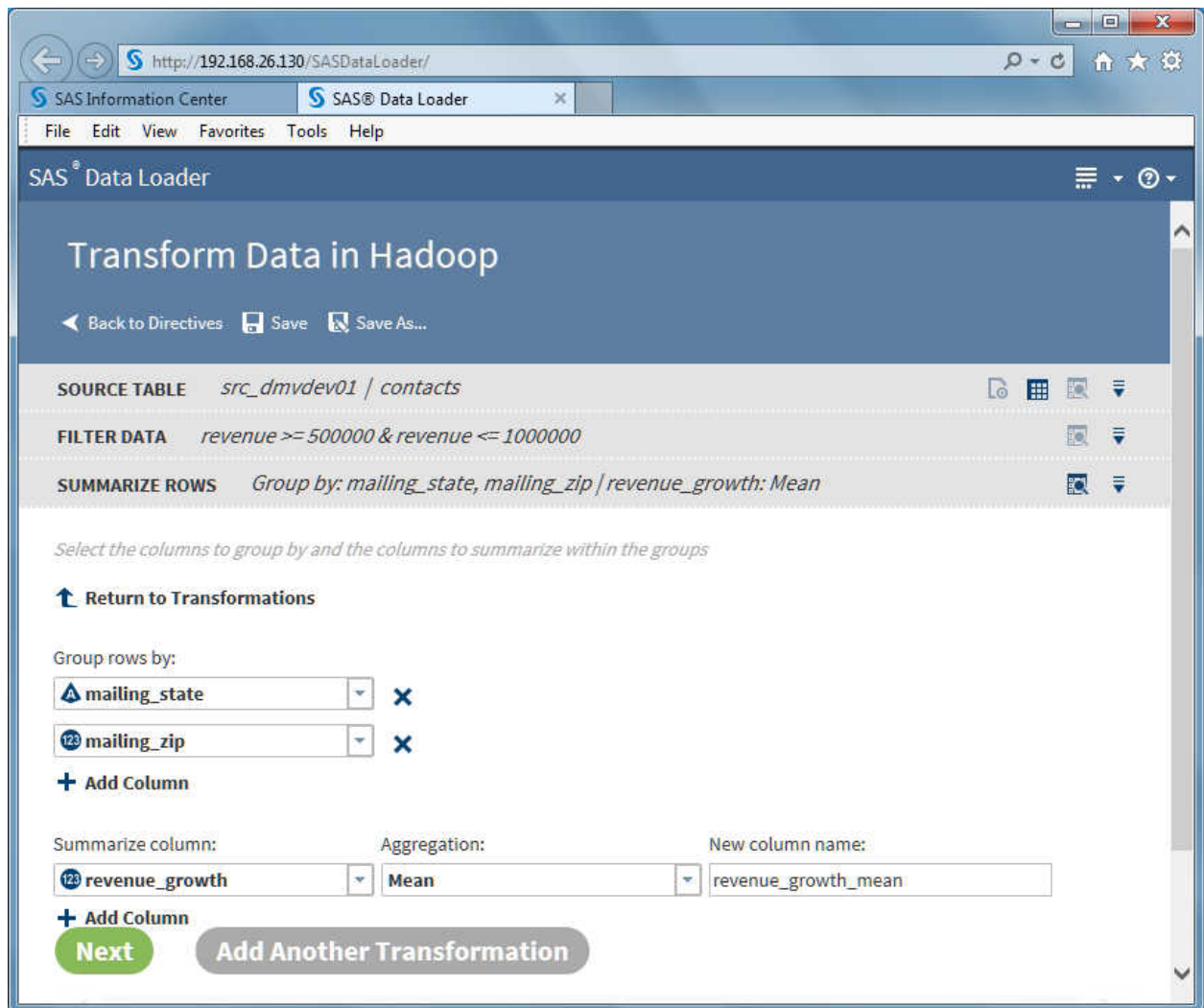
- 1 Build the job using the transformation pages in a directive, and then select a target table.
- 2 In the Result page of the directive, click **Start transforming data**.
- 3 Processing begins in the vApp, where Hadoop code is generated.
- 4 The code is submitted to the Hadoop cluster for execution. The Result page displays the Code  and Log  icons.
- 5 During execution in Hadoop, the vApp collects status messages that are sent by the Hadoop cluster.
- 6 When job execution is complete, the target table is written to disk, the job completion time is displayed, and log entries are added to the log file. Click View Results  to display the target table in the SAS Table Viewer.

In addition to executing jobs in their directives, you can also execute jobs in the directives Saved Directives and Run Status.

---

## Using the Directive Interface

Directives contain the following features.



**Back to Directives**

Click to leave the directive and display the list of directives in the SAS Data Loader for Hadoop window. Click here rather than using the Back button of the browser. You can also select **Refresh** in the browser to display the list of directives.

**Save** **Save As...**

Click at any time to save or resave your job. Open or run your job in the Saved Directives directive. If you run a job without saving it, open the Run Status directive and save it from there.

**SOURCE TABLE** *src\_dmvdev01 / contacts*

Click the transformation pages to go back to prior entries. Transformations are executed in the order in which they are defined. Changes in a prior transformation might require changes in later transformations.



Click to open user preferences, display the Table Viewer, display the data view, and delete the transformation.

### Next

Click to conclude the job by selecting a target.

**Add a new transformation**

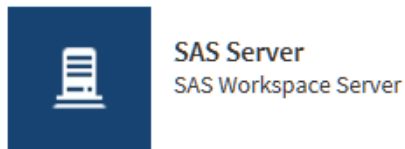
Click to open the next transformation in the sequence of the directive.

---

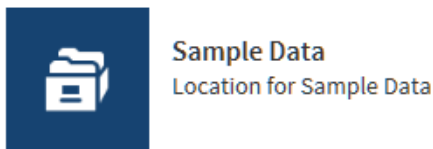
## Create and Execute a Job Using SAS Sample Data

Follow these steps to copy a small SAS sample table into Hadoop and execute a transformation on that data.

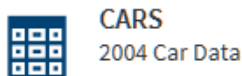
- 1 Open SAS Data Loader and click the directive Copy Data to Hadoop.
- 2 In the Source Table page, click the SAS Server data source.




- 3 Click **Sample Data**.



- 4 Click the CARS source table and click **Next**.



- 5 In the Filter page, click **Next** to filter all SAS source rows into the Hadoop target table.
- 6 In the Columns page, click **Next** to accept the existing number and arrangement of columns.
- 7 In the Target Table page, click an appropriate location for what will be a new table.
- 8 Click  **New Table...** and enter a table name such as SASCars2004.



- 9 In the Code tab, browse the generated code, and then click **Next**.
- 10 In the Result page, click **Start copying data**.
- 11 Click **View Results** to see your new table in Hadoop.



12 To transform your new table, click [Back to Directives](#).

13 In the Directives page, click **Transform Data in Hadoop**.



**Transform Data in Hadoop**  
Transform data from a Hadoop table

14 In the Source Table page, click the data source that you just used to store your new table.

15 Click your new table, and then click **Next**.



sascars2004

16 In the Transformations page, select **Summarize Rows**.



**Summarize Rows**  
Create a new row with data summarized in selected columns

17 Select group-by rows, summaries and aggregations, and then click **Next**.

Group rows by:

✕

✕

[+ Add Column](#)

Summarize column:

▼

Aggregation:

▼

New column name:

✕

▼

▼

✕

[+ Add Column](#)

18 In the Target Table page, click the data source that you use for target data, click [New Table...](#), enter a table name, and then click **Next**.

19 In the Result page, click **Start transforming data**. The job will run for a minute or so. Click View Results to see your first transformation in Hadoop using SAS Data Loader. Congratulations!

	type	model	invoice_mean	msrp_mean
1	Hybrid	Civic Hybrid 4dr man	18451	20140
2	Hybrid	Insight 2dr (gas/elect	17911	19110
3	Hybrid	Prius 4dr (gas/electri	18926	20510
4	SUV	4Runner SR5 V6	24801	27710
5	SUV	Ascender S	29977	31849
6	SUV	Aviator Ultimate	39443	42915
7	SUV	Aztek	19810	21595
8	SUV	CR-V LX	18419	19860

## Get Comfortable

### Overview

This topic introduces themes and capabilities that you will see repeatedly in SAS Data Loader for Hadoop. Look through this topic once to get comfortable, then refer back later to refresh your experience.

### About the SAS Data Loader Directives Page

In top-level web page for SAS Data Loader, you can browse and select directives. You can also select the following menus and icons:

#### Help

displays a link to the product documentation page on the SAS support website. Also displays version information, supported browsers, legal notices, and license information.

#### Configuration

opens the Configuration window, with separate panels for configuring Hadoop connections, external database connections, SAS LASR Analytic Server connections, and several categories of user preferences. These configurations are set during installation. You can reconfigure as needed to change to a different Hadoop cluster, add a new database connection, or add a connection to a grid of SAS LASR Analytic Servers.

### About Data Sources, Source Tables, and Target Tables

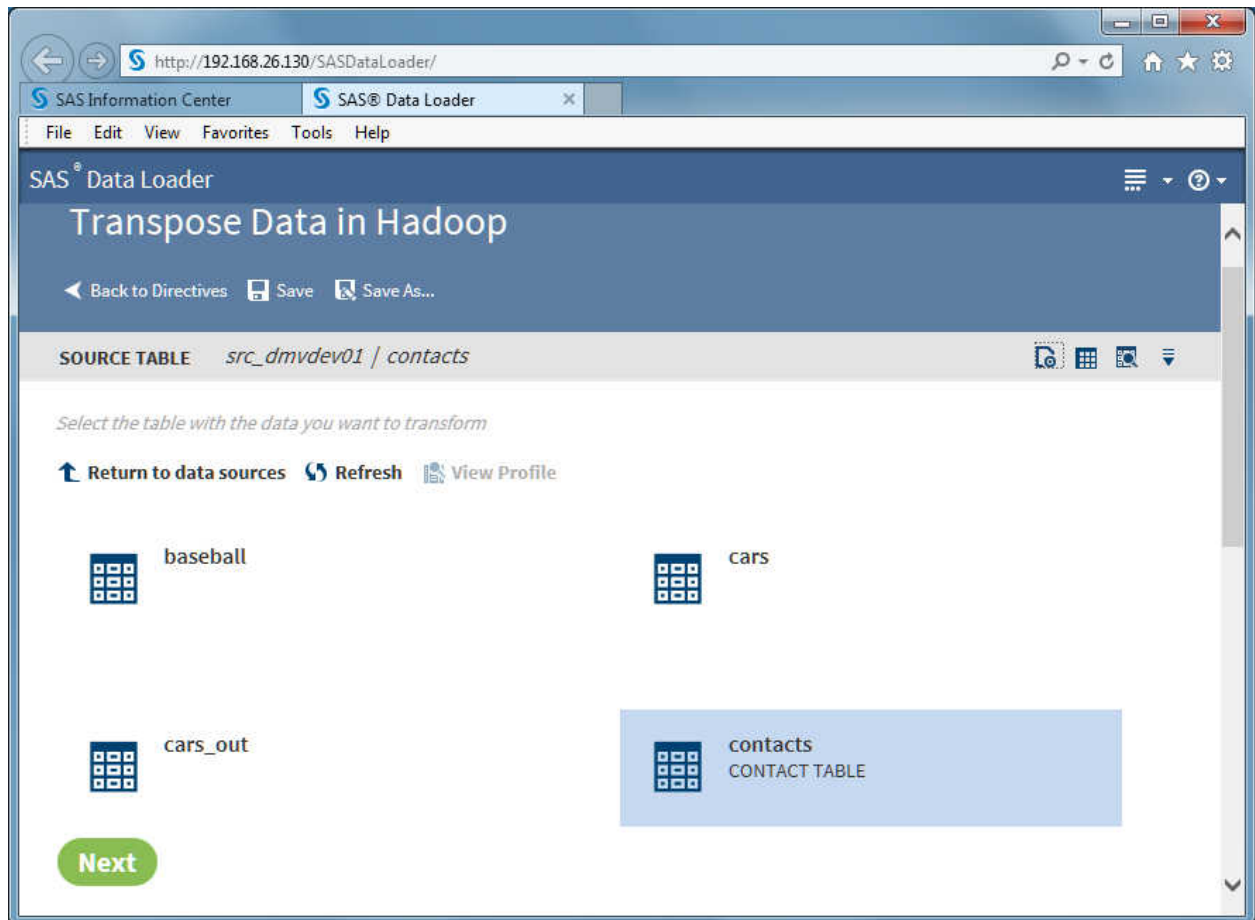
In SAS Data Loader for Hadoop, data sources contain one or more tables. Data sources are defined in the Hive database by your Hadoop administrator. If you do not see the data source or table that you need, contact your Hadoop administrator for assistance.



When you open a directive to create a job that runs in Hadoop, you select a data source and a source table that is contained within that data source. At the end of the directive, you select a data source and a target table.

To protect your data, target tables do not overwrite source tables. Target tables are not required to be new tables each time you run your job. You can overwrite target tables that you created in previous job runs.

When you run a job, data is copied from the source table into memory. As the data is processed in each task in the job, you can [view a sample](#) of the data that is produced in each task, as a means of verifying your progress.

A typical Source Table task includes a graphical view of the tables in the selected data source.



To view data and column information, to ensure that you have selected is the table that you need, click the SAS Table Viewer icon  or the View Data Sample icon .

## About the SAS Table Viewer

### How it Works

The SAS Table Viewer displays sample data and column information for a selected table. The viewer is available when you select source or target tables or when you view results or status. The SAS Table Viewer opens in a separate tab in the browser, so you can continue to reference that information while working with directives.

To open the viewer, click the **Open the selected table in the table viewer** icon



SAS Data Loader - Table Viewer

Schema name: src\_dmvdev01 Table name: cars Row limit: 100

Columns:

- ☒ make
- ☒ model
- ☒ type
- ☒ origin
- ☒ drivetrain
- ☒ msrp
- ☒ invoice
- ☒ enginesize
- ☒ cylinders
- ☒ horsepower
- ☒ mpg\_city
- ☒ mpg\_highway
- ☒ weight


Property Value

Property	Value
Index	0
Label	make
Length	13
Name	make
Type	varchar

	make	model	type	origin	drivetrain	msrp	invoice
1	Acura	MDX	SUV	Asia	All	36945	333
2	Acura	RSX Type S	Sedan	Asia	Front	23820	217
3	Acura	TSX 4dr	Sedan	Asia	Front	26990	246
4	Acura	TL 4dr	Sedan	Asia	Front	33195	302
5	Acura	3.5 RL 4dr	Sedan	Asia	Front	43755	390
6	Acura	3.5 RL w/Na	Sedan	Asia	Front	46100	411
7	Acura	NSX coupe 2	Sports	Asia	Rear	89765	799
8	Audi	A4 1.8T 4dr	Sedan	Europe	Front	25940	235
9	Audi	A4 1.8T conv	Sedan	Europe	Front	35940	325
10	Audi	A4 3.0 4dr	Sedan	Europe	Front	31840	288
11	Audi	A4 3.0 Quatt	Sedan	Europe	All	33430	303
12	Audi	A4 3.0 Quatt	Sedan	Europe	All	34480	313
13	Audi	A6 3.0 4dr	Sedan	Europe	Front	36640	331
14	Audi	A6 3.0 Quatt	Sedan	Europe	All	39640	359
15	Audi	A4 3.0 conv	Sedan	Europe	Front	42490	383
16	Audi	A4 3.0 Quatt	Sedan	Europe	All	44240	400
17	Audi	A6 2.7 Turbc	Sedan	Europe	All	42840	388
18	Audi	A6 4.2 Quatt	Sedan	Europe	All	49690	449
19	Audi	A8 L Quattr	Sedan	Europe	All	69190	647
20	Audi	S4 Quattro	Sedan	Europe	All	48040	435
21	Audi	RS 6 4dr	Sports	Europe	Front	84600	764
22	Audi	TT 1.8 conv	Sports	Europe	Front	35940	325

In the viewer, you can click a column name to display the properties of that column. You can also clear the check box next to the column name to temporarily remove that column from the sample data view.

To change the number of sample rows that are displayed, change the value of the **Row Limit** field.

To refresh the sample data after a directive has operated on that table, click the **Refresh** icon .

Column properties are defined as follows:

**Index**

Column number.

**Label**

A shortened version of the column name that can be added to the data values for that column. If a label is not assigned, then the column name is used as the label.

**Length**

The size of the table cell (or variable value) in bytes.

**Name**

Column name.

**Type**


The type of the data in the column.

For information about data types and data conversions in SAS and Hadoop, see the chapter *SAS/ACCESS Interface to Hadoop* in the document *SAS/ACCESS Interface to Relational Databases: Reference*.

**Usage Notes**

- When viewing a SQL Server table, the following numeric data types are displayed in the Columns list with a character data type: datetime (datetime\_col), money (money\_col), smallmoney (smallmoney\_col), numeric (numeric\_col), and real (real\_col).
- Viewing the source and target tables of transformations can show differences in decimal values. The source columns show no decimal values, and the target shows full double-precision values. This difference exists in the display only. In the Hadoop file system HDFS, the values are the same.

**About the Sample Data Viewer**

In directives that list tables for selection, you can click the **View a data sample** icon  to display the first 100 rows of source data, as that data has been transformed up to that point in the job. This gives you a preview of your data before you run your job in Hadoop.

Data sample:

cust_number	cust_type	cust_entity_...	cust_status	cust_since_d...	cust_since
C0000000000...	Commercial	Organization	Active	2001-12-07	Dec 7, 2001
C0000000000...	Personal	Person	Active	1996-05-18	May 18, 1996
C0000000000...	Personal	Person	Dormant	1992-06-27	Jun 27, 1992
C0000000000...	Personal	Person	Active	2005-08-21	Aug 21, 2005
C0000000000...	Personal	Person	Active	2008-04-03	Apr 3, 2008
C0000000000...	Personal	Person	Active	1991-11-12	Nov 12, 1991
C0000000000...	Personal	Person	Dormant	2005-06-06	Jun 6, 2005
C0000000000...	Commercial	Organization	Active	1993-03-07	Mar 7, 1993
C0000000000...	Commercial	Organization	Active	2012-02-26	Feb 26, 2012
C0000000000...	Personal	Person	Active	1994-06-17	Jun 17, 1994
C0000000000...	Personal	Person	Active	2006-07-08	Jul 8, 2006
C0000000000...	Personal	Person	Active	2009-10-19	Oct 19, 2009
C0000000000...	Commercial	Organization	Active	1990-01-12	Jan 12, 1990

**Next**

In the data sample, you can click **Refresh** to display the latest data or click **X** to close the data sample.

## About the Code Editor

You can edit, and save changes to the code that is generated by directives. The code editor is available in the directive's Result page. You can also edit code in the directives Run Status and Saved Directives.

The code editor is intended to be used only to implement advanced features. In normal use, there is no need to edit code. The code editor is a good way to see what will be running, but making changes can be problematic. If you make changes in the directive interface after you edit code, then your edits are lost when the code is regenerated. Also, your code edits are not reflected in the directive interface, which further complicates updates to edited code.

## 3

## Manage Data in Hadoop

<b>Overview of Data Management Directives</b>	<b>19</b>
<b>Cleanse Data in Hadoop</b>	<b>20</b>
Introduction	20
Filter Data Transformation	20
Standardization Transformation	22
Parse Data Transformation	23
Identification Analysis Transformation	24
Generate Match Codes Transformation	25
Manage Columns Transformation	26
Summarize Rows Transformation	28
About Definitions, Locales, and the SAS Quality Knowledge Base	29
About DS2 Expressions and the Advanced Editor	29
<b>Run a SAS Program</b>	<b>31</b>
Introduction	31
Example	31
<b>Query or Join Data in Hadoop</b>	<b>32</b>
Introduction	32
Example	32
<b>Sort and De-Duplicate Data in Hadoop</b>	<b>38</b>
Introduction	38
Example	39
<b>Transform Data in Hadoop</b>	<b>42</b>
Introduction	42
Example	42
About the Operators in the Filter Data Transformation	46
About the Aggregations in the Summarize Rows Transformation	50
Usage Notes	51
<b>Transpose Data in Hadoop</b>	<b>51</b>
Introduction	51
Example	51
Usage Notes	52

## Overview of Data Management Directives

The data management directives provide you with the ability to transform, join, merge a thorough set of table

## Cleanse Data in Hadoop

### Introduction



#### Cleanse Data in Hadoop

Cleanse data in Hadoop by performing data quality transforms

Use the Cleanse Data in Hadoop directive to create jobs that improve the quality of your Hadoop data. Your jobs can combine any of seven data quality transformations:



#### Filter Data

Select the rows of data to include



#### Generate Match Codes

Create match codes for selected values in the table



#### Identification Analysis

Identify the semantic data type of text in selected columns



#### Manage Columns

Select the columns to include



#### Parse Data

Select the column, Definition, and Token you want to apply, and enter a name for the new column



#### Standardize Data

Apply data standards to selected columns



#### Summarize Rows

Create a new row with data summarized in selected columns

### Filter Data Transformation

Use the Filter Data transformation at the beginning of a job to decrease the number of rows that will be processed in subsequent transformations.

Follow these steps to use the Filter Data transformation:





- 1 In the SAS Data Loader window, click **Cleanse Data in Hadoop**.



**Cleanse Data in Hadoop**  
Cleanse data in Hadoop by  
performing data quality  
transforms

- 2 In the Source Table page, click the data source that contains your source table, click the source table, and then click **Next**.

**TIP** You can examine the contents of tables using the Table Viewer (  ) or the Data Viewer (  ). For more information, see “About the SAS Table Viewer” on page 15 or “About the Sample Data Viewer” on page 17.

- 3 In the Filter Data page, you specify one or more expressions that are applied to each source row. If the expressions are true, the row is written into the target. If you specify more than one expression, you can also specify that all expressions need to be true, or only one needs to be true, to write the row into the target. In the **Include** field, you can accept the default **Rows for which all of these rules apply**, or you can select **Rows for which one or more of these rules apply**.

## Cleanse Data in Hadoop

◀ Back to Directives
💾 Save
📁 Save As...

**SOURCE TABLE**    *src\_dmvdev01 / contacts*

---

**FILTER DATA**    *primary\_state\_code = NC & revenue >= 1000000.00*

*Select the rows you want to filter.*

⬆ Return to Transformations

Include: Rows for which all of these rules apply ▼

Column:	Operator:	Value:	
<span style="border: 1px solid #ccc; padding: 2px;">primary_state_code ▼</span>	<span style="border: 1px solid #ccc; padding: 2px;">Equal To ▼</span>	<span style="border: 1px solid #ccc; padding: 2px;">NC</span>	<input type="checkbox"/> Case sensitive <span style="float: right;">✕ ?</span>
<span style="border: 1px solid #ccc; padding: 2px;">revenue ▼</span>	<span style="border: 1px solid #ccc; padding: 2px;">Greater Than or Equal To ▼</span>	<span style="border: 1px solid #ccc; padding: 2px;">1000000.00</span> ✕	<span style="float: right;">✕ ?</span>

+ Add Rule



- 4 To specify a rule (expression), click **Column** and choose a source column.
- 5 Click **Operator** and select a logical operator. The available logical operators vary by the column data type. For descriptions of the operators, see “About the Operators in the Filter Data Transformation” on page 46.
- 6 Click **Value** to specify the source value that constrains the rule. For example, in a table of business contacts, a rule could limit the companies selected to those with an annual net revenue that exceeds \$1,000,000.00
- 7 Click the Add icon **+** to specify another rule.

- 8 Click **Next** to conclude the directive and select a target table. To add another data cleansing transformation, click **Add Another Transformation**.

## Standardization Transformation

Follow these steps with your own data to learn how to use the Standardization transformation. This example creates a job that standardizes a column of state names in a table of customer data.

- 1 In the SAS Data Loader window, click **Cleanse Data in Hadoop**.
- 2 Click the schema that contains your source table. If the schema does not contain your source table, click **Return to data sources**, and open a different schema.
- 3 In the list of tables, click your source table and click **Next**.

**TIP** You can examine the contents of tables using the Table Viewer () or the Data Viewer (). For more information, see [“About the SAS Table Viewer” on page 15](#) or [“About the Sample Data Viewer” on page 17](#).

- 4 In the Cleanse Data in Hadoop window, click **Standardize Data**.
- 5 In the Standardize Data page, click **Select a Column** and select the column from the list.
- 6 Click **Select a Definition** and select the standardization definition to be applied to the selected column. Standardization definitions are available for certain character strings and numeric values. Also, standardization definitions are available for generic actions that are independent of content, such as Space Removal and Multiple Space Collapse. To learn about definitions, see [“About Definitions, Locales, and the SAS Quality Knowledge Base” on page 29](#).
- 7 Standardized values are applied to a new column in the target. You can change the default name of the new column by clicking **New column name**.
- 8 To save space or truncate long values, you can change the **Character limit** from its default value of 256.

SAS Data Loader

## Cleanse Data in Hadoop

Back to Directives Save Save As...

SOURCE TABLE *src\_dmvdev01 / customer\_1k*

STANDARDIZE DATA *cust\_street\_state\_name\_standardized*

Select the columns you want to standardize, the definition you want to apply and enter a name for the new column

Return to Transformations

Locale:  
English (United States) Select a different locale

Column:	Definition:	New Column Name:	Character limit:
cust_street_state_...	State/Province (Abbreviation)	cust_street_state_name_standardized	12

+ Add Column

Next Add Another Transformation

- 9 The standardization transformation is now completely defined. By default, the target table contains both the original source column and the new standardized column. If you would prefer to remove the source column in the target, or make other changes to target columns, add a [Manage Columns](#) transformation toward the end of your job.

Click **Next** to complete your job by selecting a target table. To continue your job, click **Add Another Transformation**.

**Note:** If your job includes multiple transformations, note that they are executed the order that you define them.

## Parse Data Transformation


Use the Parse Data transformation to extract tokens from a source column and add the token to a new column. A token is a meaningful subset of a data value that provides a basis for analysis. For example, for a column that contains phone numbers, you could extract the area code token and insert that value in a new column. You could then analyze the source table by grouping rows by area code.

Follow these steps to learn how to use the Parse Data transformation:


- 1 In the SAS Data Loader window, click **Cleanse Data in Hadoop**.
- 2 In the Source Table page, click the schema that contains your source table, click the source table itself, and then click **Next**.
- 3 In the Transformation page, click **Parse Data**.



**Parse Data**  
Select the column, Definition, and Token you want to apply, and enter a name for the new column

- 4 In the Parse Data page, click **Select a column** and select a source column from the list.
- 5 Click the **Definition** field and click the definition that you will apply to the selected column.
- 6 In the **Available tokens** list, click the token that you will copy out to a new target column.
- 7 Click the right plus arrow  to apply the token to a new column. You can change the suggested **Output Column Name**.
- 8 At this point you can choose other tokens to add to other new columns in the target table.
- 9 If you have multiple tokens, you can arrange the target columns using the up and down arrow icons.



- 10 To remove a token column, select it and click the minus arrow icon .
- 11 The Parse Data transformation is now complete. Click **Next** to conclude your job by selecting a target table. Or you can click **Add a new transformation** to continue building your job.

## Identification Analysis Transformation

Use the Identification Analysis transformation to report on the type of the content in a given column. The content types that can be detected include contact information, dates, e-mail, field names, offensive content, and phone numbers. The result of the analysis is added to a new column in the target table. You can analyze one column for multiple content types, and you can analyze multiple columns in the source table.

Follow these steps to learn how to use the Identification Analysis transformation:

- 1 In the SAS Data Loader window, click **Cleanse Data in Hadoop**.
- 2 In the Source Table page, click the schema that contains your source table, click the source table itself, and then click **Next**.
- 3 In the Transformation page, click **Identification Analysis**.







### Identification Analysis


Identify the semantic data type of text in selected columns

- 4 Click **Select a Column** and select a column for analysis.
- 5 Click **Select a Definition** and choose the content type that you want to apply to the source column.
- 6 In the **New Column Name** field, a name is suggested for the column that will be added to the target. The new column will contain the results of the identification analysis. Click in the text field for New Column Name to change the suggested column name.
- 7 To analyze another column, or to analyze the same column with a different definition, click **Add Column**.

Locale:

English (United States)  **Select a different locale**

Column:	Definition:	New Column Name:	
 contact_first_name	Contact Info	contact_first_name_id_analysis	
 last_contact_date	Date (DMY Validation - Numer...	last_contact_date_id_analysis	

 Add Column

**Next**

**Add Another Transformation**

- 8 Click **Next** to complete your job by specifying a target table. To continue your job, click **Add Another Transformation**.

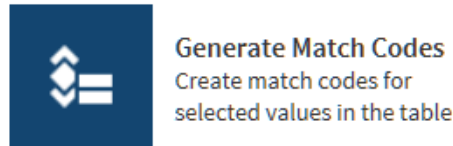
## Generate Match Codes Transformation

The Generate Match Codes transformation generates match codes for specified columns. The generated match codes are then added to new columns in the target table. The match codes are generated based on a definition and a sensitivity. The definition specifies the type of the content in the column. The sensitivity determines the degree of exactitude that is required in order for two data values to be declared a match. Higher sensitivity values specify that data values must be more similar to be declared a match. Lower sensitivity values enable matching with less similarity. The level of sensitivity is reflected in the length and complexity of the match codes.

Match codes can be used to find columns that contain similar data. For example, you can generate match codes for name and address columns, and then compare the match codes to detect duplicates.

Follow these steps to use the Generate Match Codes transformation:

- 1 In the SAS Data Loader window, click **Cleanse Data in Hadoop**.
- 2 In the Source Table page, click the schema that contains your source table, click the source table itself, and then click **Next**.
- 3 In the Transformation page, click **Generate Match Codes**.



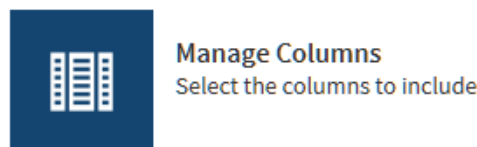
- 4 Click **Select a Column** and then click the column for which you want to generate match codes.
- 5 Click **Select a Definition** and then click the definition that you want to use to generate match codes.
- 6 To change the default sensitivity value, click the **Sensitivity** field and select a new value. Lower sensitivity numbers give you more matches (identical match codes) and perhaps more matching errors. Higher sensitivity numbers produce the same match code only when data values are nearly identical.
- 7 This completes the definition of the Generate Match Codes transformation. To generate match codes for the same column using a different definition, or to generate match codes for a different column, click **Add Another Transformation**. Otherwise, click **Next** to conclude your job and select a target table.

## Manage Columns Transformation

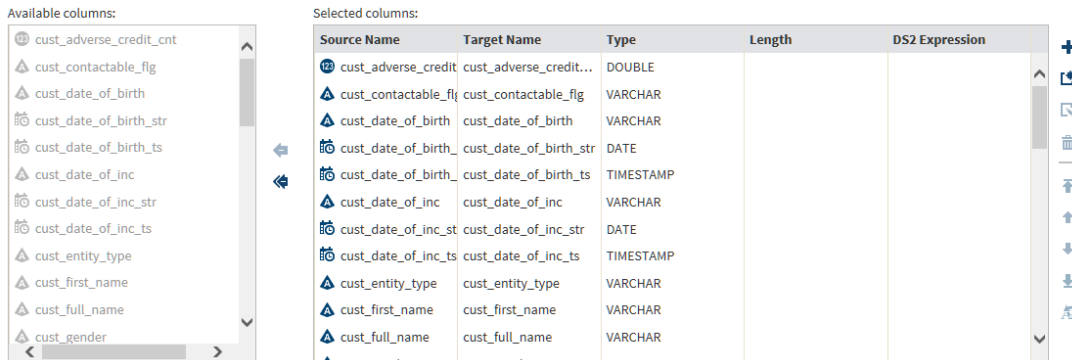
Use the Manage Columns transformation to remove, reorder, and rename source columns. You can also add new columns. The new columns contain generated values of a specified length and type. The values are generated by a DS2 expression that you supply, based on the values in each row. To learn more about DS2, see the *SAS 9.4 DS2 Language Reference*.

Follow these steps to learn how to use the Manage Columns transformation:

- 1 In the SAS Data Loader window, click **Cleanse Data in Hadoop**.
- 2 In the Source Table page, click the schema that contains your source table, click the source table itself, and then click **Next**.
- 3 In the Transformation page, click **Manage Columns**.



- 4 Click **Manage Columns** to open a new job or to add a transformation to an existing job.



- 5 In the Manage Columns page, columns are listed in order of appearance. The top column is the first or leftmost column.

Note the arrow icons between **Available columns** and **Selected columns**. To remove a column from the target, click the column name on the right and click the top arrow. To move all columns out of the target, click the double-arrow icon. After you remove a column, arrows will appear so that you can move columns from Available to Selected.

Initially, all columns are selected for the target table, including all of the new that you added in prior transformations.

- 6 Now locate the icons on the right side of **Selected columns**. These icons provide the following functions:
- Add a new column and enter a DS2 expression for that column without using the Advanced Editor.
  - Add a new column and specify a DS2 expression using the Advanced Editor.
  - Edit the selected column using the Advanced Editor to modify its DS2 expression.
  - Remove the selected column from the target table. Removed columns appear in **Available columns**.
  - Move the selected column to the first column position in the target (leftmost.)
  - Move the selected column one position to the left in the target.
  - Move the selected column one position to the right.
  - Move the selected column to the last column position in the target (rightmost.)
  - Change the name of the selected target column.
- 7 If you want to add or paste a DS2 expression into an existing column, click the DS2 Expression field for that column and proceed. Any source data in that column will be replaced by the results of the DS2 expression.
- 8 If you want to use the Advanced Editor to define a DS2 expression, click and see [“About DS2 Expressions and the Advanced Editor”](#) on page 29.

- 9 When your target columns are ready, click **Next** to conclude your job by specifying a target, or click **Add Another Transformation**.

## Summarize Rows Transformation

Use the Summarize Rows transformation to add summarized numeric values to your target table. To generate summaries, you first group rows by one or more columns. Then you select the columns that you want to summarize for each group and subgroup. The method of summarization is known as an aggregation. The number of aggregations depends on the column data type. Numeric columns have 13 available aggregations.

Follow these steps to learn how to use the Summarize Rows transformation:

- 1 In the SAS Data Loader window, click **Cleanse Data in Hadoop**.
- 2 In the Source Table page, click the schema that contains your source table, click the source table itself, and then click **Next**.
- 3 In the Transformation page, click **Summarize Rows**.



### Summarize Rows

Create a new row with data summarized  
in selected columns

- 4 Click the **Group rows by** field and choose the first column that you want to use to group rows for summarization. In the target table, rows with the same values in the selected column appear together, along with their summary values in new columns.
- 5 To further subset the initial set of groups, and to generate a second set of summary values, click **Add Column**. Select a second column. Add additional groups as needed.
- 6 Click **Summarize column** and select the first numeric column that you want to summarize.
- 7 Click **Aggregation** and select the aggregation that you would like to provide for the selected column.
- 8 To change the suggested name for the new column that will contain the aggregation values for each group, click **New Column Name**.
- 9 To add a second aggregation, click **Add Column**.



Select the columns to group by and the columns to summarize within the groups

[Return to Transformations](#)

Group rows by:

primary\_state\_code

primary\_zip

Add Column

Summarize column:

Aggregation:

New column name:

net\_income

Mean

net\_income\_mean

income\_growth

Mean

income\_growth\_mean

Add Column

Next

Add Another Transformation

- 10** Your Summarize Row transformation is now complete. Click **Next** to conclude your job and select a target table. To continue your job, click **Add Another Transformation**.

## About Definitions, Locales, and the SAS Quality Knowledge Base

In the SAS data quality software, *definitions* specify the usage of terms within a *locale*. A locale consists of a language and a region. The region is frequently a country. The default locale is English (United States). You can read the locale name as “the English language, as it is used in the United States.” You can change the current locale using **Select a locale**.



Locale:

English (United States) **Select a different locale**



In the selected locale, definitions are provided for each type of transformation. Transformations apply definitions to columns in source tables. For the standardization transformation, you can for example apply the Phone definition to a column of phone numbers. The transformation converts all of the values in the specified column into the phone number format that is specified in the definition.

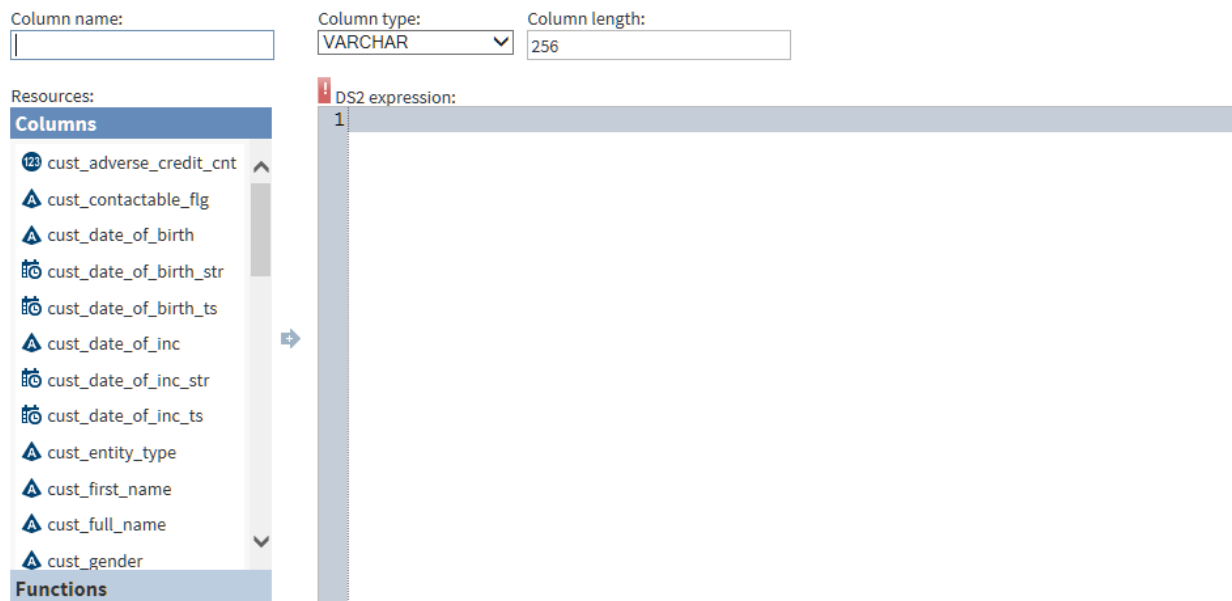
## About DS2 Expressions and the Advanced Editor

In the Manage Columns transformation, you can add new columns and specify DS2 expressions for those columns. When you run your job, the DS2 expression is evaluated for each row and the result is added to the new column.

When you add a new column, you can enter or paste a DS2 expression directly into the **DS2 Expression** column (click ) , or you can add your DS2 expression in the Advanced Editor (click ) . In either case, your expression uses DS2 expression syntax (and not SAS expression syntax.) For information about DS2 expressions, refer to the *SAS 9.4 DS2 Language Reference*.

Follow these steps to learn more about the Advanced Editor:

- 1 In the Manage Columns transformation, click  to add a new column and open the Advanced Editor. Note that you can also select an existing column and click  to replace the data in that column with the results of a DS2 expression.



Column name:

Column type:

Column length:

Resources:

**Columns**

- cust\_adverse\_credit\_cnt
- cust\_contactable\_flg
- cust\_date\_of\_birth
- cust\_date\_of\_birth\_str
- cust\_date\_of\_birth\_ts
- cust\_date\_of\_inc
- cust\_date\_of\_inc\_str
- cust\_date\_of\_inc\_ts
- cust\_entity\_type
- cust\_first\_name
- cust\_full\_name
- cust\_gender

**Functions**

DS2 expression:

1

- 2 Enter a name for the new column, a column data type, and the length of the column in bytes (if applicable.) The **Column type** is the data type of the result of your DS2 expression.
- 3 Define your DS2 expression using the columns and functions in the **Resources** list.

**TIP** When you select a function, help is displayed for that function at the bottom of the **Resources** list.

- 4 When your DS2 expression is complete, click **Save** to return to the Manage Columns page. If you defined a new column for your DS2 expression, the new column appears at the bottom of the **Selected columns** list.

## Run a SAS Program

### Introduction




**Run a SAS Program**  
Run in-database data quality SAS programs

Use the directive Run a SAS Program to execute SAS programs in Hadoop. These programs are written with ultra-efficient SAS DS2 language elements. The DS2 elements combine the power of SAS with threaded SQL calls. All of this execution takes place on the Hadoop cluster.

For information about programming with DS2, see the *SAS 9.4 DS2 Language Reference*.

### Example

Follow these steps to use the Run a SAS Program transformation:

- 1 Develop and test a SAS DS2 program. The program needs to explicitly define any and all sources and targets. The directive Run a SAS Program supports SAS DS2 programs only.
- 2 In the SAS Data Loader window, click **Run a SAS Program**.
- 3 Copy the text of your SAS DS2 program, and then paste that program into the Code page.
- 4 Edit your SAS DS2 program as needed. When the program is ready for execution in Hadoop, click **Next**.
- 5 Click **Start SAS program**.
- 6 In the Code page, right-click and select **Paste**.  
**Note:** In the pop-up menu in the Code page, the following options are invalid: Navigate out of code (backward), Navigate out of code (forward), and Syntax Help.
- 7 Verify that your entire program is now present in the text editor, edit the program as needed, and then click **Next**.
- 8 In the Result page, click **Start SAS Program**. The directive runs and generates a selectable Log icon . The final status of the directive is portrayed by an icon in the Result banner.  
You can monitor the execution of your SAS program in the “[Run Status](#)” directive.
- 9 Click **Save** or **Save As** to store your directive in your local Shared Folder.

---

## Query or Join Data in Hadoop

### Introduction



#### Query or Join Data in Hadoop

Query a table, or join data from multiple tables

Use queries to group rows based on the values in one or more columns and then summarize selected numeric columns. The summary data appears in new columns in the target table.


Use joins to combine source tables. The join is based on a comparison of values in “join-on” columns that are selected for each of the source tables. The result of the join depends on matching values in the join-on columns, and on the selected type of the join. Four types of joins are available: inner, left, right, and full.

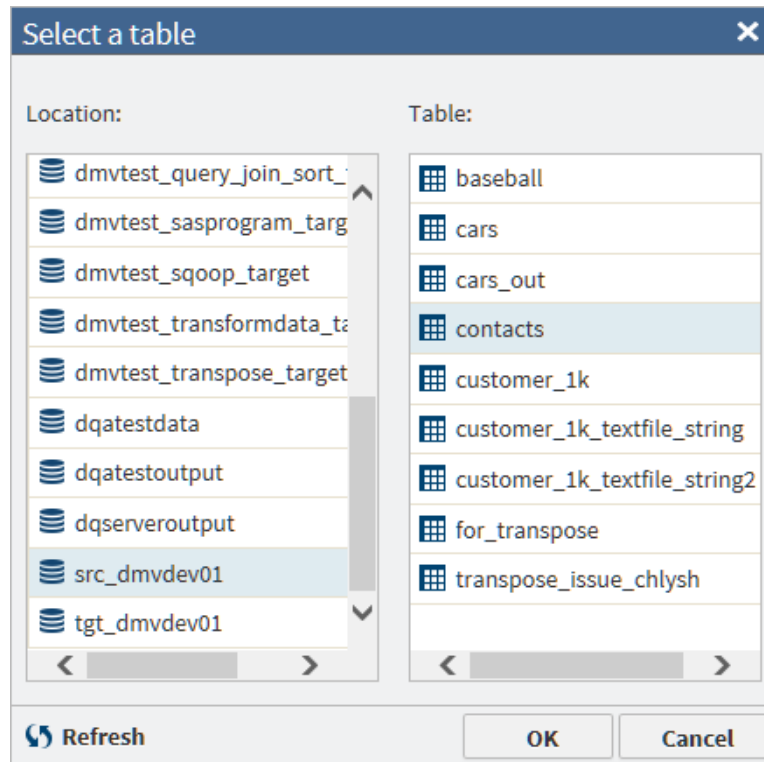
The Query or Join Data in Hadoop directive enables you to create jobs that combine multiple joins and queries, and then customize the target table to remove unwanted rows and columns, remove duplicate rows, and rearrange columns. Before you execute the job you can edit the Hive SQL code and paste in additional Hive SQL code. The process of the directive is defined as follows:


- Select a source table.
- Join tables to the initial table as needed.
- Define queries that group columns and aggregate numeric values, again as needed.
- For jobs that do not include queries, use rules to filter unwanted rows from the target. (Queries require all rows.)
- For join-only jobs, select, arrange, and rename target columns.
- For join-only jobs, apply Hive SQL expressions in new or existing target columns.
- Sort target rows based on specified target columns.

### Example

Follow these steps to use the Query or Join Data In Hadoop directive.

- 1 In the SAS Data Loader page, click **Query or Join Data in Hadoop**.
- 2 In the Query page, click the browse icon .
- 3 In the Select a Table window, scroll through the **Location** list and click a schema. Then click a source table in the **Table** list, and then click **OK**.



- 4 If your job includes no joins, click **Next** to open the Summarize Rows page.
- 5 To join your source table with other tables, click **Add Join**, and then click **Next**.
- 6 In the **Join** row, click the browse icon  and select the table for the join.
- 7 As needed, click the **Join** field and select a join type other than the default join type **Inner**.

#### Inner

The inner join finds matching values in the join-on columns and writes one row to the target. The target row contains all columns from both source tables. A row from either source table is not written to the target if it contains a null value in the join-on column. A row is also not written to the target if the value in the join-on column does not match a value in the join-on column in the other source table.

#### Left

The left or left-full join writes to the target all rows from the left table of the join statement. If a match does not exist between the join-on columns, null values are written to the target for the columns of the right table in the join.

#### Right



The right or right-full join reverses the definition of the left join. All rows from the right table appear in the target. If no values match between the join-on columns, then null values are written to the target for the columns of the table on the left side of the join statement.

#### Full

The full join combines the left and right joins. If a match exists between the join-on columns, then a single row is written to the target to represent those two source rows. If the left or right table has a value in the join-on

column that does not match, then the data for that row is written to the target and null values are written into the columns from the other source table.

- 8 In the **Join-on** row, click the left join-on column and select a replacement for the default column, as needed.

Join on:  src\_dmvdev01.contacts.contact\_last\_name ▼ =  src\_dmvdev01.customer\_1k.cust\_last\_name ▼ +

**Note:** The left and right designations in the join-on statement define the output that is generated by the available left join and right join.

- 9 Click the right join-on column to select a replacement for the column, as needed.
- 10 To add more join columns, click the Add icon + at the end of the **Join-on** row. When you add a second pair of join-on columns, a match between the source tables consists of a match in the first pair of join-on values *and* a match between the second pair of join-on values.
- 11 To join a third table to the joined table that unites the two source tables, click **Add join**.



## Query or Join Data in Hadoop


◀ Back to Directives  Save  Save As...

**JOIN** Inner Join: contacts.contact\_full\_name = customer\_1k.cust\_full\_name, contacts.contact\_dob = customer\_1k.cust\_date\_of\_birth

Choose a table to query, or multiple tables to join and the columns to join on

Base table:  src\_dmvdev01.contacts ...

Join:  Inner Join  src\_dmvdev01.customer\_1k ... ✕

Join on:  src\_dmvdev01.contacts.contact\_full\_name ▼ =  src\_dmvdev01.customer\_1k.cust\_full\_name ▼ +

and  src\_dmvdev01.contacts.contact\_dob ▼ =  src\_dmvdev01.customer\_1k.cust\_date\_of\_birth ▼ + ✕

+ Add Join

Next

- 12 Click **Next** and wait a moment while the application assembles in memory the names of the joined columns.
- 13 In the Summarize Rows page, if you do not need to summarize, click **Next**.
- 14 To add summarizations, click the **Group rows by** field, and then click the column that you want to use as the primary grouping in your target table. For example, if you are querying a table of product sales data, then you could group rows by the product type column.

**Note:**

- If your job includes joins, note that the **Group rows by** list includes all columns from your source tables.

- If you intend to paste a Hive SQL query into this directive, then you can click **Next** two times, to bypass the pages for summaries and filters and reach the Code page.

**15** To subset the first group with a second group, and to generate a second set of aggregations, click **Add column**.

**16** To generate multiple aggregations, you can add additional groups. The additional groups will appear in the target table as nested subgroups. Each group that you define will receive its own aggregations.

To add a group, click **Add Column**, and then repeat the previous step to select a different column than the first group. In a table of product sales data, you could choose a second group by selecting the column `product_code`.

**17** In **Summarize columns**, select the first numeric column that you want to aggregate.

**18** In **Aggregations**, select one of the following:

Count

specifies the number of rows that contain values in each group.

Count Distinct

specifies the number of rows that contain distinct (or unique) values in each group.

Max

specifies the largest value in each group.

Min

specifies the smallest value in each group.

Sum

specifies the total of the values in each group.

**19** In **New column name**, either accept the default name of the aggregation column, or click to specify a new name.

**20** To add an aggregation, click **Add Column**.

## Query or Join Data in Hadoop

◀ Back to Directives  Save  Save As...

**JOIN** Inner Join: contacts.contact\_full\_name = customer\_1k.cust\_number, contacts.contact\_dob = customer\_

**SUMMARIZE ROWS** Group by: mailing\_state, mailing\_zip / total\_employees: Max / revenue: Max

Select the columns to group by and the columns to summarize within the groups

Group rows by:

 src\_dmvdev01.contacts.m... 

 src\_dmvdev01.contacts.m... 

+ Add Column

Summarize column:

Aggregation:

New column name:


 src\_dmvdev01.contacts.to... 

Max




total\_employees\_max 

 src\_dmvdev01.contacts.re... 

Max

revenue\_max 

+ Add Column


- 21 When the aggregations are complete, click **Next**.
- 22 In the Filter Data page, all rows are included in the target by default. To accept this default, click **Next**.
- 23 If your job includes joins but no summarizations, then you can select **No duplicate rows** to remove duplicate rows from the target.
- 24 The field **Include rows where** applies when you specify multiple rules. The default value **Rows for which all of these rules apply** writes rows to the target only if all of the rules apply. When you select **Rows for which any of these rules apply**, rows are written to the target only if one or more of the rules apply.
- 25 Click **Select a column** and choose the column for the rule.
- 26 Click **Operator** to specify a logical operator for your rule. The logical operators that are available for your rules depend on the data type of your column. Columns can be numeric , character , or datetime . To learn about the available logical operators, see [Table 3.1 on page 46](#).
- 27 To add another rule, click **Add Rule**. When your filter rules are complete, click **Next**.



**FILTER ROWS** *last\_contact\_date < 1/1/2014 | salary < 60000*

Select the rows you want to filter.

☐ All rows ☐ Specify rows

☒ No duplicate rows 

Include: **Rows for which any of these rules apply** 

Column:	Operator:	Value:
 src_dmvdev01.contacts.la... 	Before 	1/1/2014   
 src_dmvdev01.contacts.sa... 	Less Than 	60000  













 Add Rule

**Next**

**28** Click **Next** to open the Columns page. Use the Columns page to select, order, and rename the columns that will be written into the target table. Also use the Columns page to apply Hive SQL expressions to new or existing columns.



The Columns page is available only if your job does not contain summaries. If your job does contain summaries, then click **Next** to display the Sort page, and then click **Next** again to open the Target Table page.

**29** Use the Columns page to do the following:

- Select and order the columns in the target, using the arrow icons to select all , select one , remove one  (or ) , and remove all .
- Replace the suggested column names as needed by clicking in the **Target Name** column.
- Reorder columns by clicking a column and clicking  (move to first column),  (move column left one position),  (move column right one position), and .
- Add new columns for Hive SQL expressions. Click the Add icon , and then specify the column name.
- Add new column and open the Advanced Editor to develop a Hive SQL expression. Click .
- Add a Hive SQL expression to an existing column using the Advanced Editor. Click .

**30** Follow these steps to use the Advanced Editor to generate a Hive SQL expression:

**31** Click **Next** to close the Columns page and open the Target Table page.

- 32 In the Target Table page, to learn about the contents of a table, click the table and click the Table Viewer icon .
- 33 To write your target data to an existing table, click that table and click next. Any and all existing data will be replaced.
- 34 To save data to a new target table, click  **New Table...**, enter a table name in the New Table window, and click **OK**.  
  
The names of tables must meet the naming conventions of SAS and Hadoop.
- 35 To display your target data as a temporary view, click ☒ **Save as a View**.  
  
Saving as a view displays your target data in the [Sample Data Viewer](#) without saving the results to a target table on disk.  
  
When your target selection is complete, click **Next** to open the Code page
- 36 In the Code page, click **Edit HiveQL Code** to edit the generated code. Click **Reset Code** to restore the original generated code. Click **Next** to open the Result page.  
  
**Note:** Edit your Hive SQL code with care. The code in the editor is the exact code that will be executed by your job, regardless of previous selections.
- 37 In the Result page, you can review the previous pages by clicking on the gray action bars at the top of the window. and making changes.
- 38 Click **Save** or **Save As** to save your job.
- 39 Click **Start querying data** to execute your directive. To monitor the progress of your job, see the [“Run Status”](#) directive.

---

## Sort and De-Duplicate Data in Hadoop

### Introduction



**Sort and De-Duplicate ...**  
Query, sort, or de-duplicate  
the data in an existing  
Hadoop table

Use the Sort and De-Duplicate Data in Hadoop directive to create jobs that include some or all of the following steps:

- 1 If needed, group rows based on selected columns and then summarize selected numeric columns for each group.
- 2 If not summarizing, specify the removal of duplicate rows.
- 3 Filter rows into the target table by applying rules to selected columns.

- 4 Remove, reposition, and rename the columns in the target table. Add columns for HiveQL expressions as needed.
- 5 Sort target rows by selecting one or more columns for ascending or descending values.

## Example



Follow these steps to use the Sort and De-Duplicate directive:

- 1 Open SAS Data Loader, as described in [Chapter 2, “Get Started,”](#) on page 5.
- 2 In the Source Table page, click a schema, and then, when it appears, click a source table.
- 3 In the Summarize Rows page, if you do not want to generate summary values for groups of rows, click **Next**. Otherwise, click **Group rows by** and select a column. To generate nested groups with additional summary values, click **Add Column**.

Group rows by:

 mailing_state	▼	✕
 mailing_zip	▼	✕
<b>+ Add Column</b>		

- 4 Click **Summarize column** and select a column that will be used to generate summary values for each group.
- 5 Click **Aggregation** and select the summary type.
- 6 Click **New column name** to change the default column name. The new column will contain a summary value for each group.
- 7 Click **Add Column** to specify a second summary and new column.

Summarize column:	Aggregation:	New column name:			
 salary	▼	Sum	▼	salary_sum	✕
 salary	▼	Count	▼	salary_count	✕
<b>+ Add Column</b>					





- 8 Click **Next** to filter rows into the target table. By default, all source rows are filtered into the target. If you do not need filters, click **Next** again.
- 9 In the Filter Rows page, click **Specify rows**. You can also select **No duplicate rows** if your job does not include summaries.
- 10 Filter rules are simple expressions that evaluate the values in specified columns. If the expression evaluates to true, then the source row can be written to the target. You can define multiple rules. If you have multiple rules, and if all rules must evaluate to true to write a row to the target, then accept the default value for the **Include** field.

If rows are to be written to the target when one or more rules are true, then select the value **Rows for which any of these rules apply**. Click **Add rule** as needed.

☐ All rows ☒ Specify rows

☐ No duplicate rows 

Include: **Rows for which all of these rules apply** 

Column:	Operator:	Value:
 <b>contact_since_ts</b> 	<b>On or After</b> 	1/1/2010 
 <b>last_contact_ts</b> 	<b>On or After</b> 	1/1/2014  

**+ Add Rule**


- 11** Click **Next** to display the Manage Columns page. If you defined summaries in the Summarize Rows page, or if you do not need to remove, reorder, rename, or add HiveQL expressions to the columns in the target table, click **Next** again.
- 12** In the Manage Columns page, click **Specify Columns**.
- 13** To remove columns from the target, click the left and right arrow icons between **Available columns** and **Selected columns**.






- 14** To reorder the selected columns, use the vertical arrows.



- 15** To rename selected columns, click and enter or paste a new name in the **Target Name** column.
- 16** You can generate values for the target table using HiveQL expressions. The values of those expressions can go into existing columns (replacing the source data in those columns,) or the values can go into new columns. You can create the HiveQL expressions by entering or pasting them directly into the **Hive Expression** column. You can also use the Advanced Editor to generate HiveQL expressions for new or existing target columns. To

generate a new column that will contain a HiveQL expression that you create in the Advanced Editor, click .

- 17 In the Advanced Editor, in the **Column Name** field, enter a name for a new column or rename an existing column.
- 18 Click the column names and functions in **Resources** to build your expression. When you select a function, syntax help is displayed at the bottom of **Resources**.
- 19 To save your expression and return to the Columns page, click **Save**. To save and create another new column and expression, click **Save and New**. In the Columns page, the new columns are displayed at the bottom of **Selected Columns**.

 Save  Save and New  Cancel

Column name:

Column type:

Column length:

Resources:  

Columns

Functions

All Functions

-  
!  
!=  
%  
&  
&&  
\*  
/

(no functions selected)

Hive expression:

1 (table0.employee\_growth / table0.total\_employees) \* 100

- 20 When the target columns are complete, click **Next**. If you have not defined any summaries in the Summarize Rows page, then the Sort page enables you to group rows based on ascending or descending values in specified columns. If you defined summaries, then the Target Table page enables you to select an existing target table or create a new target table.
- 21 In the Sort page, select one or more columns and a sort order of **Ascending** or **Descending** for each column. Sorts are nested in the target in the order in which they are defined. Click **Next** to open the Target Table page.

*Choose columns to sort by*

 mailing\_state

Ascending



 primary\_zip

Ascending



+ Add Column

- 22 In the Target Table page, select a location for the target table. When the table list appears, either select an existing target or click **New Table**. To generate a temporary table that is not saved to disk, select **Save as a View**. Click **Next**.
- 23 In the Code page, review and edit the generated HiveQL code. Note that your edits might not be reflected in the preceding pages of the directive. Click **Next**.
- 24 In the Result page, click **Save** or **Save As** to save your job and list it in Saved Directives. Click **Start querying data** to run your job.

---

## Transform Data in Hadoop

### Introduction



**Transform Data in Hadoop**  
Transform data from a Hadoop table

Use the Transform Data in Hadoop directive to filter data, manage columns, and summarize data in one or more Hadoop source tables.

### Example

The following example depicts the process of creating and running a directive that contains several transformations. The example opens a source table of customer information, selects columns for the target, and applies two filters.

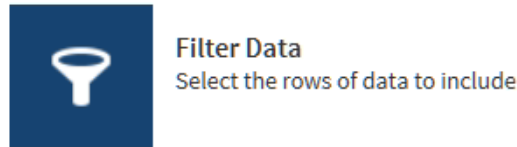
- 1 Open SAS Data Loader, as described in in [Chapter 2, “Get Started,”](#) on page 5.
- 2 In the SAS Data Loader page, click **Transform Data in Hadoop**.
- 3 In the Source Table page, click the schema that contains the source table that you will transform. When the tables appear, select the source table and click **Next**.
- 4 In the Transformation page, click a transformation:
  - Click **Filter Data** to define rules that include only desired data in the target.
  - Click **Manage Columns** to manage the columns in your target table. You can select source columns, reorder columns, and rename columns. You can also add or repurpose target columns to store the results of DS2 expression. An advanced editor is provided to assist with the development of DS2 expressions.

**Note:** To apply HiveQL expressions rather than DS2 expressions, see the Manage Columns transformation in the Query or Join Data in Hadoop directive.

- Click **Summarize Columns** to group rows based on the values in one or more columns. For each group, you can generate summary aggregations from selected numeric columns.

Your job can consist of one or more transformations. Multiple transformations are executed in the order in which you define them. A logical order for all three transformations is filter data, manage columns, and summarize columns.

5 Click **Filter Data**.

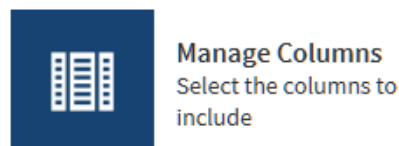


- 6 Select the columns that you will use to filter the rows that will be written into the target table. For example, in a table of customer information, you could limit the data in your target to customers with incomes between \$40,000 and \$80,000. This filter requires two rules, and both rules must be true in order for the source row to be written to the target.
- 7 In the Filter Data page, accept the default value for the **Include** field: **Rows for which all of these rules apply**.
- 8 Select columns, operators, and values to define rules.

Column:	Operator:	Value:		
123 cust_gross_annual_income	Greater Than or Equal To	40000	X	?
123 cust_gross_annual_income	Less Than or Equal To	80000	X	?

The operators that are available depend on the type of the column. To learn about available operators, see [“About the Operators in the Filter Data Transformation” on page 46](#).

- 9 At this point, you could end a job that consists solely of a Filter Data transformation. You would click **Next** to select a target table and run your job. Instead, to see the other two available transformations, click **Add Another Transformation**.
- 10 In the Transformation page, click **Manage Columns**.



- 11 Determine the columns that you want to see in your target table. In a table of customer data, you could choose columns for full name, gross annual income, net worth, number of adverse credit events, and state code. These columns include those that will be used in a Summarize transformation.
- 12 In the Manage Columns page, use the left and right arrow icons to click and move columns into and out of the **Selected Columns** list. Columns are listed


vertically, with the first or leftmost column at the top, and the last or rightmost column at the bottom.

- 13** Use the vertical arrow icons to change the position of the columns.

The screenshot shows the 'Manage Columns' interface. On the left, under 'Available columns:', there is a list of columns: cust\_number, cust\_type, cust\_entity\_type, cust\_status, cust\_since\_date, cust\_since\_date\_str, cust\_since\_datetime, cust\_since\_datetime\_str, and cust\_tax\_id. On the right, under 'Selected columns:', there is a table with the following data:

Source Name	Target Name	Type	Length	DS2 Expression
cust_last_name	cust_last_name	VARCHAR		
cust_street_state_co	cust_street_state_code	VARCHAR		
cust_last_contact_d	cust_last_contact_d...	DATE		
cust_gross_annual_i	cust_gross_annual_i...	DOUBLE		
cust_net_worth_amc	cust_net_worth_am...	DOUBLE		
cust_adverse_credit	cust_adverse_credit...	DOUBLE		

Vertical arrow icons are visible on the right side of the 'Selected columns' table to allow reordering.

- 14** To rename columns, click and enter or paste the new name in **Table Name**.
- 15** To replace existing column data with data that is generated by a DS2 expression, click a selected column and click the **DS2 Expression** column. Enter or paste the DS2 expression.
- 16** To add a new column, and to use the Advanced Editor to generate a DS2 expression for that column, click .
- 17** To use the Advanced Editor, enter a **Column Name**, and then apply DS2 functions to specified target columns. When you select a column, syntax help appears at the bottom of **Resources**. When your DS2 expression is complete, select **Save** or **Save New** to return to the Manage Columns transformation. The new column appears at the bottom of **Selected Columns**.



Column name:  Column type:  Column length:

Resources:

**Columns**

**Functions**

- All Functions
- Aggregate
- Arithmetic
- Array
- Bitwise Logical Operation
- Character
- Character String Matchin
- Combinatorial
- Date and Time

(no functions selected)

DS2 expression:

```
1 cust_net_annual_income / cust_num_relations
```

- 18 Click **Add a new transformation**, and then, in the Transformation page, click **Summarize**.
- 19 In the Summarize Rows page, click **Group rows by** to specify a column whose values will be used to group rows. You can specify additional columns that will form subgroups. Each group and subgroup will receive a value in each aggregation column.
- 20 Click **Select a column** to specify a summarization, and then click and select an aggregation. to learn about the available aggregations, see [“About the Aggregations in the Summarize Rows Transformation” on page 50](#).
- 21 Click **New column name** and enter or paste replacement names for the aggregation columns.


Group rows by:


+ Add Column

Summarize column:  Aggregation:  New column name:

- 22 When your summaries are complete, click **Next** to conclude your job.

**23** In the Target Table page, select the schema that contains or will contain your target table.

**24** Click  **New Table...** to create a new table, or click an existing table that will be overwritten by your job.










**TIP** If you select a table and View Profile  **View Profile** is enabled, you can click that icon to display a profile report for that table.








**25** Click **Next** to display the Result page. In the Result page, click **Save** or **Save As** to store your job in your shared folder. If you want to run your job now, click **Start transforming data**. Otherwise, you can run your job later from “Saved Directives”.









## About the Operators in the Filter Data Transformation










The following table describes filter operators by the data type of the selected column.


**Table 3.1** Logical Operators in the Filter Transformation

Operator	Source ColumnData Types	Description and Example
Equal To	<p>The Equal To operator is available for use with all source data types, which include the following:</p> <p>Character </p> <p>Numeric </p> <p>Datetime </p>	<p>The source value is accepted and its row is written to the target table only when the source value exactly matches the comparator.</p> <p>Character values can be case-sensitive. Blank spaces are included in the comparison.</p> <p>Datetime values in the comparator use the SAS format DATETIME(w.p).</p> <p>Gender Equal To Male</p> <p>PrefCustomer Equal To 1</p> <p>SaleDate Equal To 5/1/2014</p>
Not Equal To	  	<p>Accepts the source row when the column value is anything other than the comparator.</p> <p>Region Not Equal To Europe</p> <p>NumChildren Not Equal To 0</p> <p>SaleDate Not Equal To 11/25/2013</p>
Null	  	<p>Accepts the source row when the column value is NULL or if no source value is present.</p> <p>CreditScore Null</p> <p>AnnualIncome Null</p>

Operator	Source ColumnData Types	Description and Example
Not Null	  	<p>Accepts the source row when the column value is present and when the value is not NULL.</p> <pre>PostalCode Not Null PhoneNumber Not Null</pre>
In	 	<p>Accepts the source row when the column value is included in its entirety within the comparator. The comparator consists of a list of constant values. The list consists of a vertical list of individual entries, without commas. Blank spaces are interpreted literally. Case sensitivity can be enabled.</p> <pre>CarManuf In BMW VW Benz WaistSize In 32 34 36 38</pre>
Not In	 	<p>Accepts the source row when the column value is not included anywhere within the comparator's list of constant values.</p> <pre>City Not In New York Chicago Los Angeles WaistSize Not In 32 34 36 38</pre>

Operator	Source ColumnData Types	Description and Example
Like	 	<p>Accepts the source row when the column value matches the result of an expression in the comparator. The source value and the comparator are compared on a character-by-character basis. For character columns, case sensitivity can be enabled.</p> <p>Use the pattern-matching character % to indicate any string of characters. Use the underscore character _ to indicate any single character in that position.</p> <p>Note that trailing blank characters are written to the target table when using % at the end of the comparator.</p> <p>Use the word <code>escape</code> to include literal instances of % and _ in the comparator.</p> <p><code>SalesRegion Like NorthAmer%</code>  <code>AnnualSales Like 199_</code>  <code>CustSatisfaction Like 100 escape %</code></p>
Not Like	 	<p>Accepts the source row when the column value does not match the result of an expression in the comparator. The source value and the comparator are compared on a character-by-character basis. For character columns, case sensitivity can be enabled. Pattern-matching characters % and _ and <code>escape</code> are valid as described for the Like operator.</p> <p><code>Sports Not Like %ball</code>  <code>FootballFieldLength Not Like 100%</code></p>
Contains	 	<p>Accepts the source row when the column value is found within the character string of the comparator. Case sensitivity can be enabled.</p> <p><code>Address Contains IL</code>  <code>LicenseNumber Contains 7227</code></p>
Not Contains	 	<p>Accepts the source row when the column value is not found within the character string of the comparator. Case sensitivity can be enabled.</p> <p><code>Month Not Contains OctNovDec</code>  <code>SalesMonthly Not Contains 0</code></p>

Operator	Source ColumnData Types	Description and Example
Between	 	Accepts the source row when the column value or date is between the two values or dates in the comparator, but is not equal to either.  GradeAverage Between 87.5 93  DailySales Between December 20, 2014 December 27, 2014
Greater Than		Accepts the source row when the column value is greater than the value of the comparator.  AnnualSales GreaterThan 100000
Greater Than Or Equal To		Accepts the source row when the column value is equal to the comparator or greater than the comparator.  CarsInFamily Greater Than or Equal To 3
Less Than		Accepts the source row when the column value is less than the value of the comparator.  GamerAge Less Than 30
Less Than Or Equal To		Accepts the source row when the column value is equal to the value of the comparator, or less than the value of the comparator.  SalesYear Less Than Or Equal To 2010
After		Accepts the source row when the column date is later than the date in the comparator.  HomePurchaseDate After January 1, 2013
Before		Accepts the source row when the column date is earlier than the date in the comparator.  BirthDate Before March 17, 1980
On Or After		Accepts the source row when the column date is later than, or the same date as, the date in the comparator.  DailySales On Or After January 1, 2014

Operator	Source ColumnData Types	Description and Example
On Or Before		Accepts the source row when the column date is earlier than, or the same date as, the date in the comparator.  DailySales On Or Before December 31, 2013

## About the Aggregations in the Summarize Rows Transformation

The aggregations that are available in the Summarize Rows transformation are defined as follows:

### Count

the number of rows in the group that contain valid values.

### Count Distinct

the number of unique values in the column for each group.

### Corrected Sum of Squares

measures variability or dispersion around the mean. To learn more about this (and other) statistical summaries, see the *Introduction to Statistical Modeling with SAS/STAT Software*.

### Covariance

measures the strength of the correlation of the values in the group. A positive value indicates that values move in the same direction within the group. A negative value indicates that values move in opposite or random directions.

### Max

the maximum value in the column for each group.

### Mean

the calculated center value between the maximum and minimum values in the group.

### Min

the minimum value in the group.

### Number of Missing Values

the number of rows in the group that contain a blank or NULL value.

### Range

the difference between the lowest and highest values in the group.

### Standard Deviation

measures the degree of variance, or the degree in which the values in the group deviate from the mean. A small value indicates little deviation. The standard deviation is the square root of the Variance.

### Standard Error

measures the applicability or accuracy of the mean as it applies to the values in the group. A small value indicates that the mean is a more accurate reflection of the values in the group.

- Sum
  - adds the values in the group
- Variance
  - the average of the squared differences from the mean, which measure diversity in the group

## Usage Notes

See the usage note “[Changing the Default Maximum Length for SAS Character Columns](#)” on page 109.

---

# Transpose Data in Hadoop

## Introduction



**Transpose Data in Hadoop**  
Transpose data from a Hadoop table


Use the Transpose Data in Hadoop directive to transpose one or more columns in a source table into rows in a target table. The columns in the target are the values of a specified column in the source. For example, you could specify that the columns of the target be taken from the values of a source table column that contains customer ID numbers. Each unique customer ID value in the source becomes a separate column in the target.

You do not have to transpose all of the columns in the source. You can select source columns that will be copied directly to the target.

This directive contains embedded help that includes examples of the transposed data.

## Example

Follow these steps to use the Transpose Data in Hadoop directive.

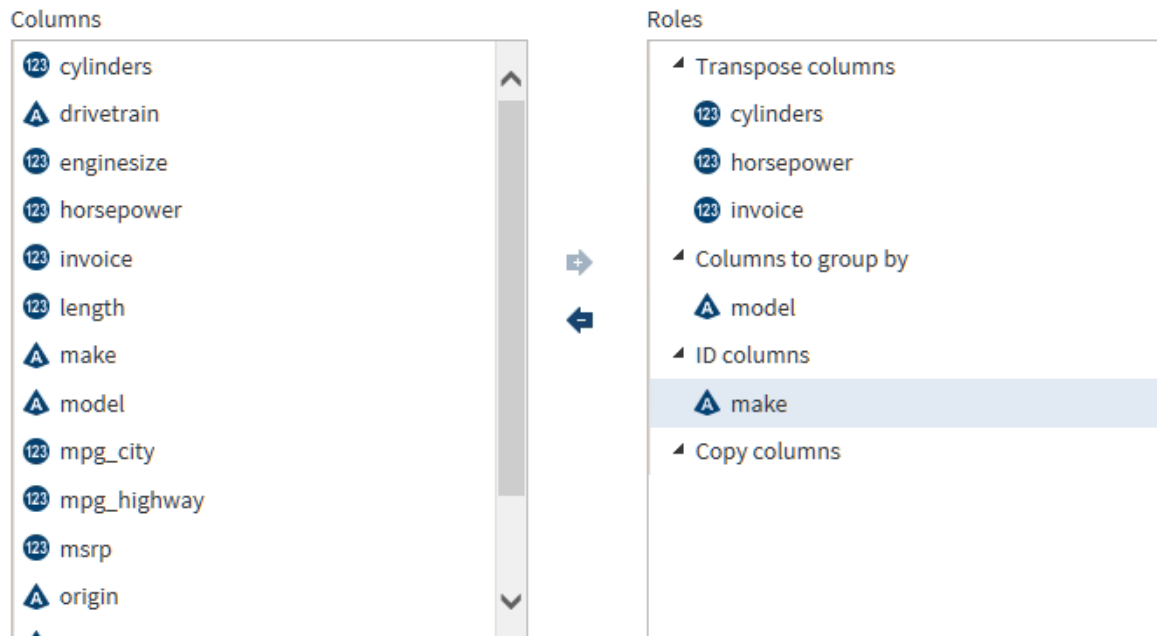
- 1 Open SAS Data Loader and click **Transpose Data in Hadoop**.
- 2 In the Source Data page, click the data source that contains your source table, click the source table, and then click the Table Viewer .

Examine the source table to determine the roles for the columns.

**Note:** Valid source table selections must have names that contain no more than 32 characters. Longer table names cause transpose jobs to fail.

- 3 In the Transpose Data page, click the required **Transpose data**, click the columns that you want to see as rows, and click the right arrow. If you transpose multiple columns, then you can arrange them in **Roles** using the up and down arrows.

- 4 Click the required **Columns to group by**, click an available column, and then click the right arrow. The group-by column becomes the leftmost column. Each row in that column receives a set of values from the transposed columns.
- 5 As needed, click **ID column**, click an available column, and then click the right arrow. The values of the ID column become column names in the target.



- 6 To copy a column from the source to the target, select **Copy column**, select an available column, and click the right arrow. The copied column will be positioned as the last, or rightmost, column.

## Usage Notes

See the usage note “[Changing the Default Maximum Length for SAS Character Columns](#)” on page 109.



## 4

## Profile Data in Hadoop

<b>Overview of Profile Directives</b> .....	<b>53</b>
<b>Profile Data</b> .....	<b>55</b>
Introduction .....	55
Table Name Length Requirement .....	55
Configure Profile Jobs .....	55
Create a Profile .....	56
<b>Saved Profile Reports</b> .....	<b>62</b>
Introduction .....	62
Open Saved Profile Reports .....	62

---

### Overview of Profile Directives

The profile directives enable you to generate and view reports for one or more Hadoop tables. The reports display sample data, column information, and measurements of data quality. You create profile reports with the Profile Data directive and use the Saved Profile Reports directive to access and manage profile reports.

Here's an example of a profile report:

## SAS® Data Loader - Profile Reports



Test

[Go to Profile Report List](#)
[Show Outline](#)
[Show Notes](#)
[Add Note...](#)

Report Version:

Jan 28, 2015, 10:46:00 AM



Overview &gt; tgt\_dmvdev01.transpose\_singlecol

Count: 7

## Data Quality Metrics

Column	#	Unique (n)	Unique (%)	Pattern (n)	Pattern (%)	Null (n)	Null (%)	Blank (n)	Blank (%)
<a href="#">tester1</a>	1	7	100	2	28	0	0	0	0
<a href="#">_name_</a>	2	1	14	1	14	0	0	0	0
<a href="#">col1</a>	3	3	50	1	14	1	14	0	0

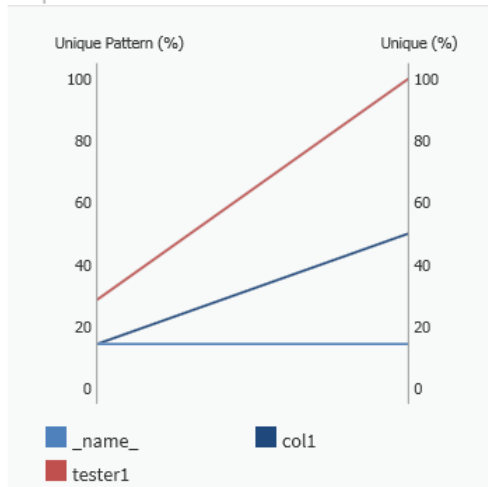
\* indicates data not available or not applicable for this column.

## Descriptive Measures

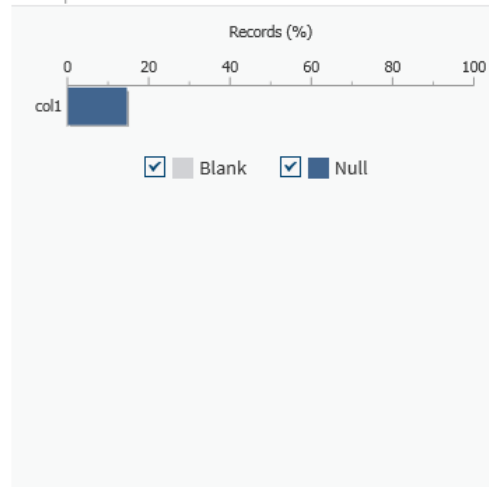
## Metadata Measures

## Charts

Uniqueness



Incompleteness



## Profile Data

### Introduction



#### Profile Data

Generate a profile report of the data in a table

Use the Profile Data directive to generate profile reports for one or more tables. You can select a subset of the columns that you want to include in the profile report.

### Table Name Length Requirement

Hive tables have a maximum table name length of 132 characters. Many of the SAS Data Loader directives can create tables with names that exceed the SAS table name length limit of 32 characters. The tables that you submit for profiling in the Profile Data directive must conform to the 32-character name length limit. Table names that exceed 32 characters generate error messages.

### Configure Profile Jobs

Follow these steps to view and edit the properties for profile jobs:

**Note:** These are advanced settings that are not normally changed. In normal operation, the default values are sufficient.

- 1 On the SAS Data Loader page, click  and select **Configuration**. The Configuration dialog box is displayed:

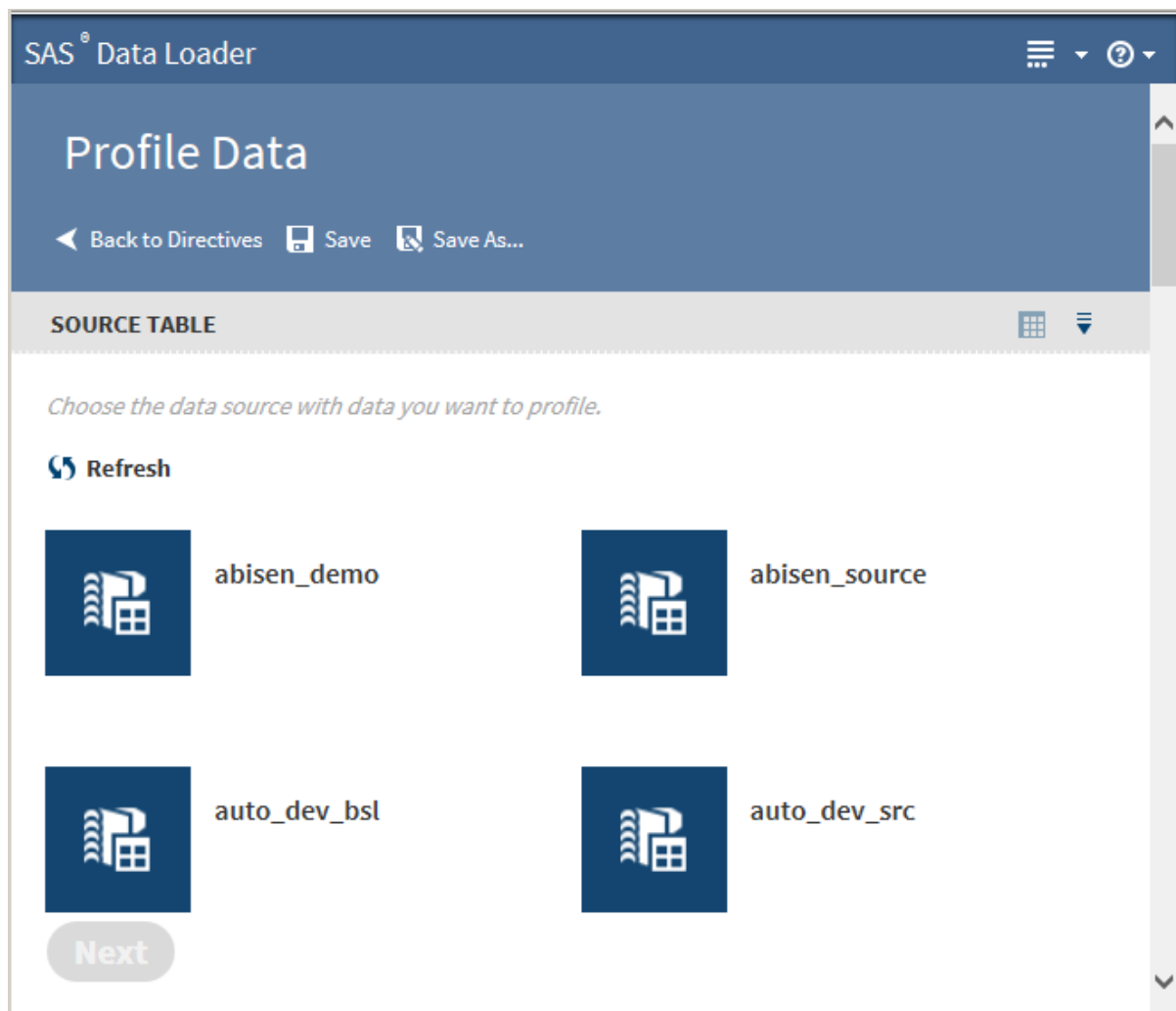
In the Configuration dialog box, click **Profiles**.

- 2 For **Stop processing a column...**, select the check box to stop processing a column if the number of unique values is greater than or equal to the number that you enter in the field.
- 3 For **Maximum number of frequency...**, select the check box and in the field enter the maximum number of frequency distribution values to save. If there are more frequency distribution values than this number, the less-frequent values are combined into an Other frequency distribution.
- 4 For **Number of outlier values to save**, select the check box and in the field enter the maximum number of outlier values to save.
- 5 Click **OK** to close the window and save your changes.

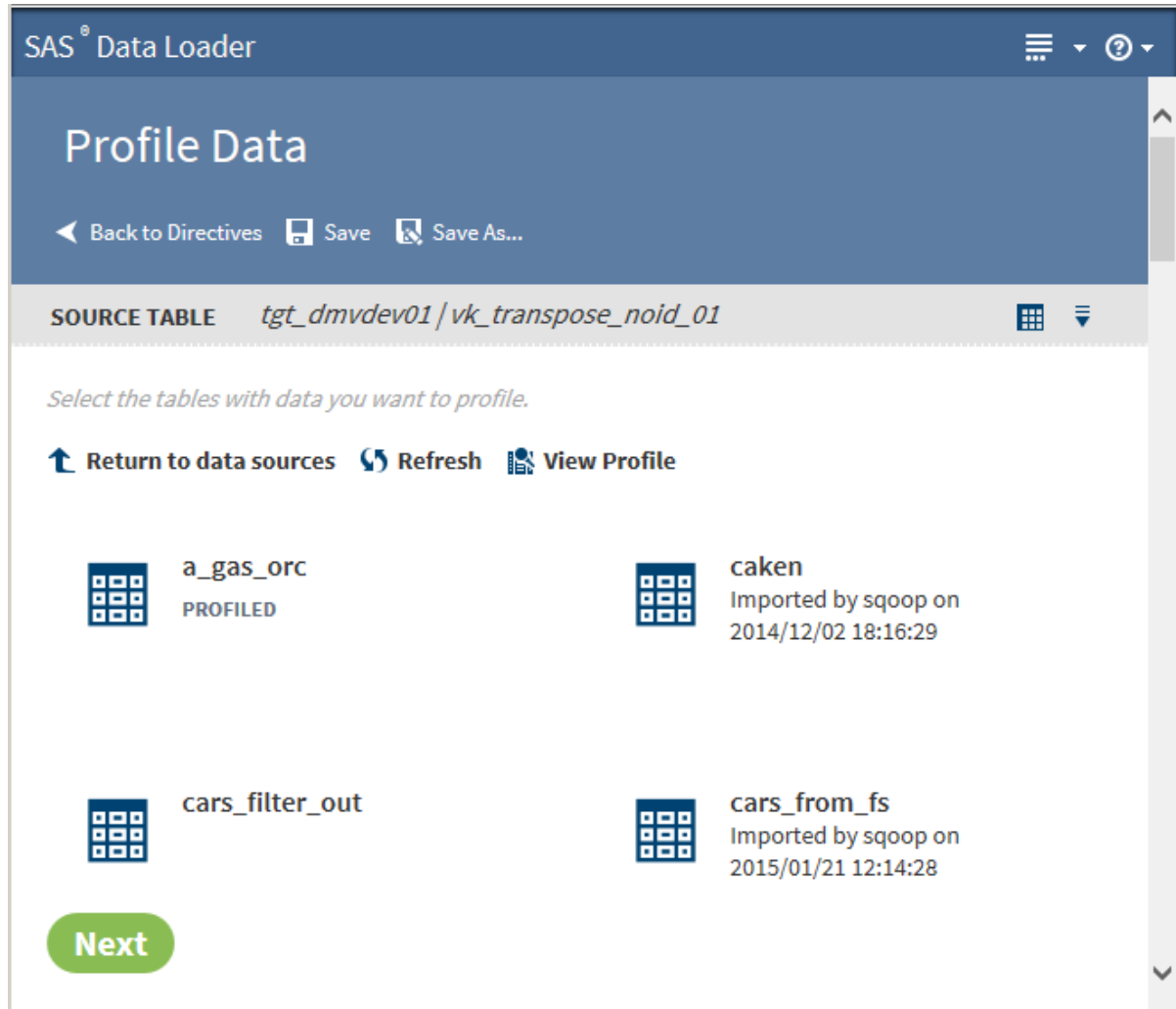
## Create a Profile

To create a profile:


- 1 On the SAS Data Loader page, click the Profile Data directive. The Source Table page is displayed:



2 Click a data source to display its tables:



3 Select the table or tables for the profile report.

**TIP** To view sample data from a table, select the table, and then click  in the Source Table header to display the [SAS Table Viewer](#).

**TIP** If a profile already exists for a table, PROFILED appears beneath the table name. You can view the existing profile by selecting the table and clicking **View Profile**.

**TIP** Profile results are also available in the Filter Transformation. If a table has been profiled, an ellipsis button appears next to the filter value selection. Click that button to view profile results while building your filters.

Click **Next**. The Columns page is displayed:

SAS® Data Loader

# Profile Data

[Back to Directives](#)
[Save](#)
[Save As...](#)

**SOURCE TABLE** *tgt\_dmvdev01 / transpose\_singlecol, vk\_transpose\_...*

**COLUMNS** *8 of 9 columns*

Select the columns you want to profile.

- ▶ *tgt\_dmvdev01.transpose\_singlecol* (3 of 3 columns)
- ▼ *tgt\_dmvdev01.vk\_transpose\_noid\_01* (5 of 6 columns)

**Available columns:**





- byfield
- \_name\_
- col1
- col2
- col3
- col4

**Selected columns:**

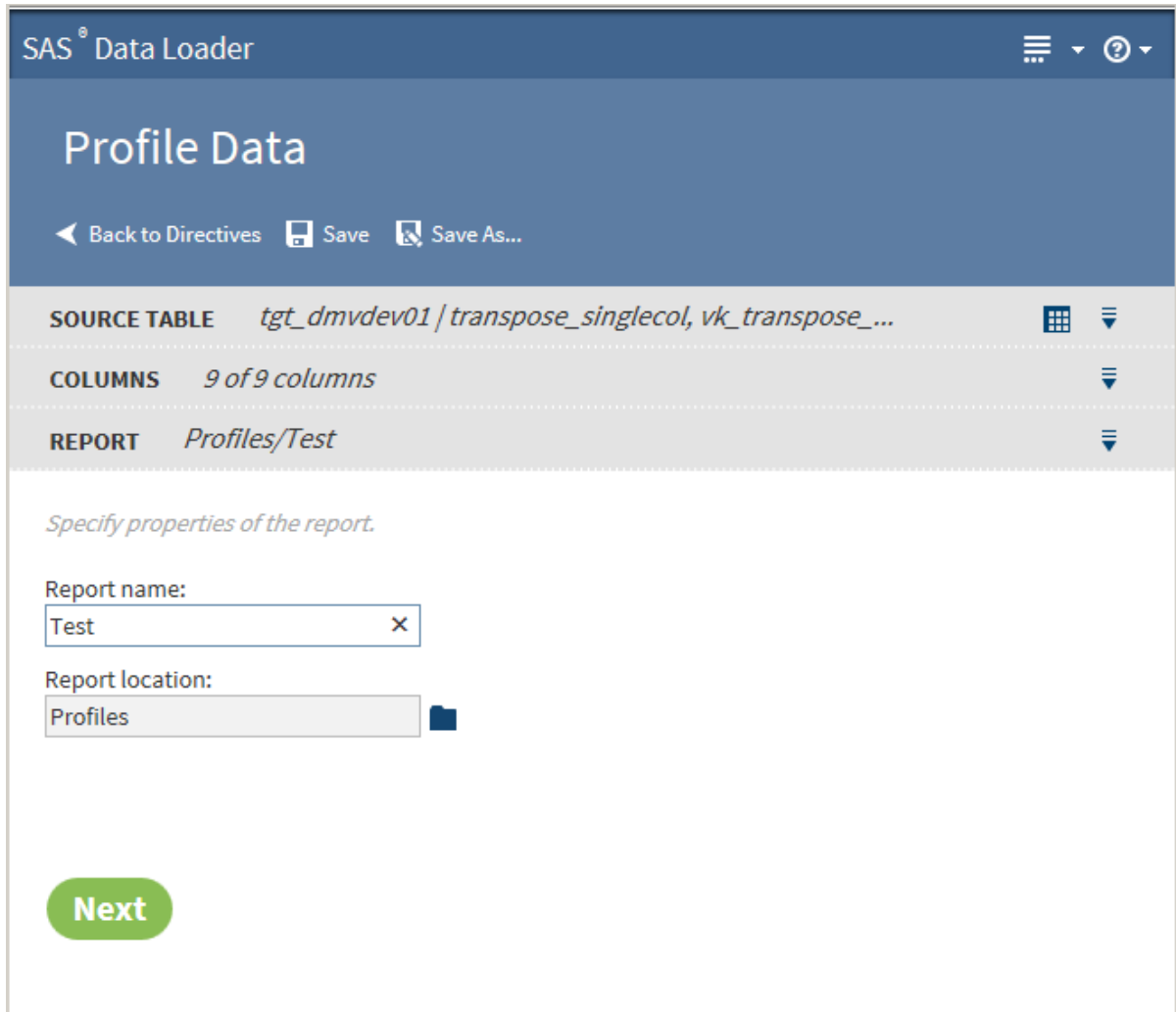
- byfield
- \_name\_
- col1
- col2
- col3

Next


- The Columns page displays the total number of columns that are to be processed in the profile report. If you selected more than one table for your report, the tables are listed by name. Click ▶ next to the tables to display the columns that are included in the profile report.

- 5 The column names in the **Selected columns** pane appear in the report. Select an individual column name and click  or  to move the column name between the **Available columns** pane and the **Selected columns** pane until the correct list of names appears in the **Selected columns** pane. Click  or  to move all column names at once.

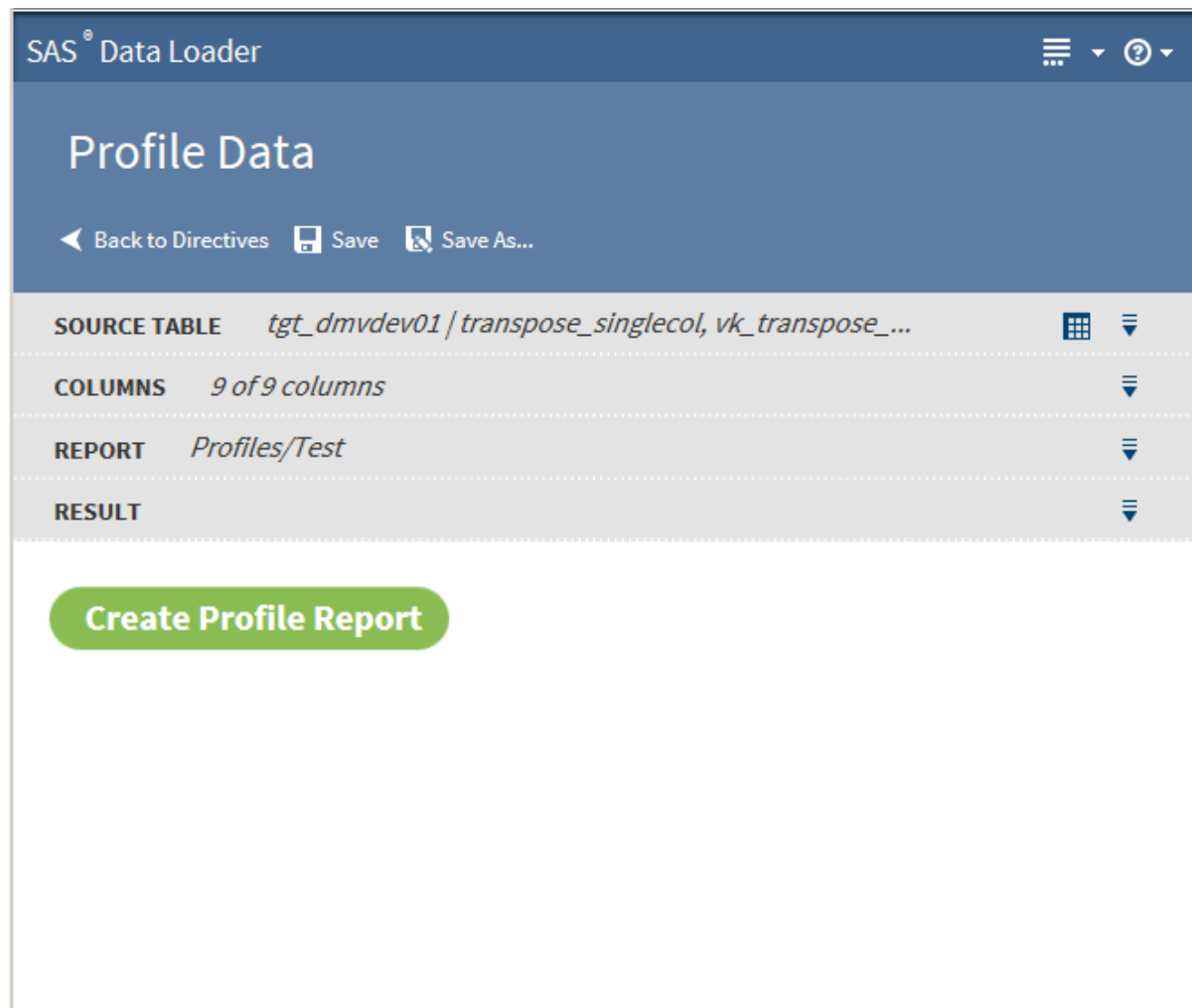
When the column selection is complete, click **Next**. The Report page is displayed:



The screenshot shows the 'Profile Data' configuration window in SAS Data Loader. The window has a dark blue header with the title 'SAS® Data Loader' and a menu icon. Below the header, the title 'Profile Data' is displayed. A navigation bar contains 'Back to Directives', 'Save', and 'Save As...' buttons. The main area shows configuration details: 'SOURCE TABLE' is 'tgt\_dmvdev01 / transpose\_singlecol, vk\_transpose...', 'COLUMNS' is '9 of 9 columns', and 'REPORT' is 'Profiles/Test'. Below this, a section titled 'Specify properties of the report.' contains two fields: 'Report name:' with a text box containing 'Test' and a close button, and 'Report location:' with a text box containing 'Profiles' and a folder icon. At the bottom left, there is a large green 'Next' button.

- 6 In the Report page, enter a name for the profile report in the **Report name** field. Click  next to the **Report location** field to change the storage location of the profile report.

After specifying a name and location, click **Next**. The Result page is displayed:



- 7 Click **Create Profile Report**. After successfully creating the profile report, a screen similar to the following is displayed:



SAS Data Loader

## Profile Data

[Back to Directives](#)
[Save](#)
[Save As...](#)

<b>SOURCE TABLE</b>	<i>tgt_dmvdev01 / transpose_singlecol, vk_transpose_...</i>		
<b>COLUMNS</b>	<i>9 of 9 columns</i>		
<b>REPORT</b>	<i>Profiles/Test</i>		
<b>RESULT</b>	<i>Successfully profiled data</i>		

Started January 28, 2015 at 10:46:16 AM EST  
 Completed January 28, 2015 at 10:51:05 AM EST

**View Profile Report**
**Log**
**Code**

**Create Profile Report**

The following actions are available:

**View Profile Report**

enables you to view the Profile Report. See [“Saved Profile Reports” on page 62](#) for more information about the profile report.

**Log**

displays the SAS log that is generated during the creation of the profile.

**Code**

displays the SAS code that generates the profile.

## Saved Profile Reports

### Introduction



#### Saved Profile Reports

Explore previously generated profile reports

Use the Saved Profile Reports directive to view the results of previously executed data profiles and to create notes about the results. The profiles are created with the Profile Data directive. The profile reports and notes are stored as XML documents on the file system. Saved Profile Reports displays these XML files in a readable format.




### Open Saved Profile Reports

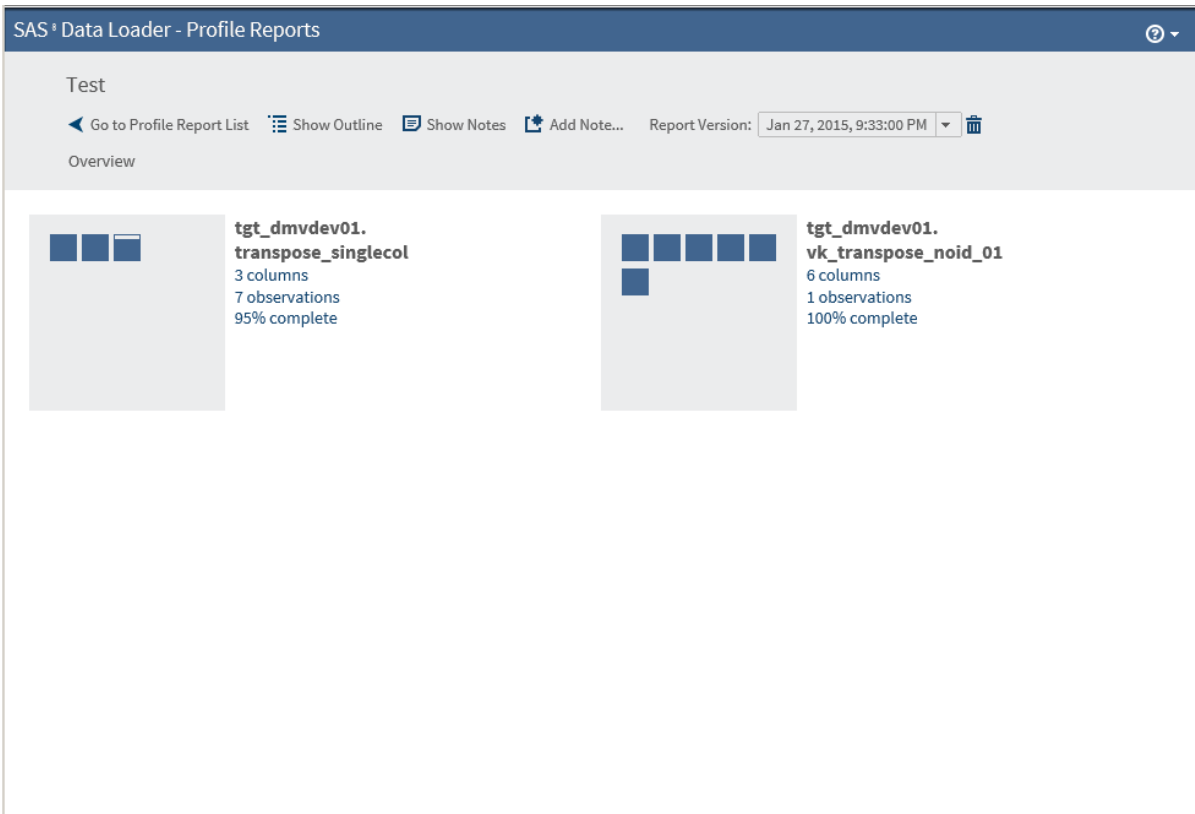
To open a saved profile report:


- 1 In the SAS Data Loader page, click the Saved Profile Reports directive to open a new browser tab. The Select a Profile Report page is displayed on the new tab:

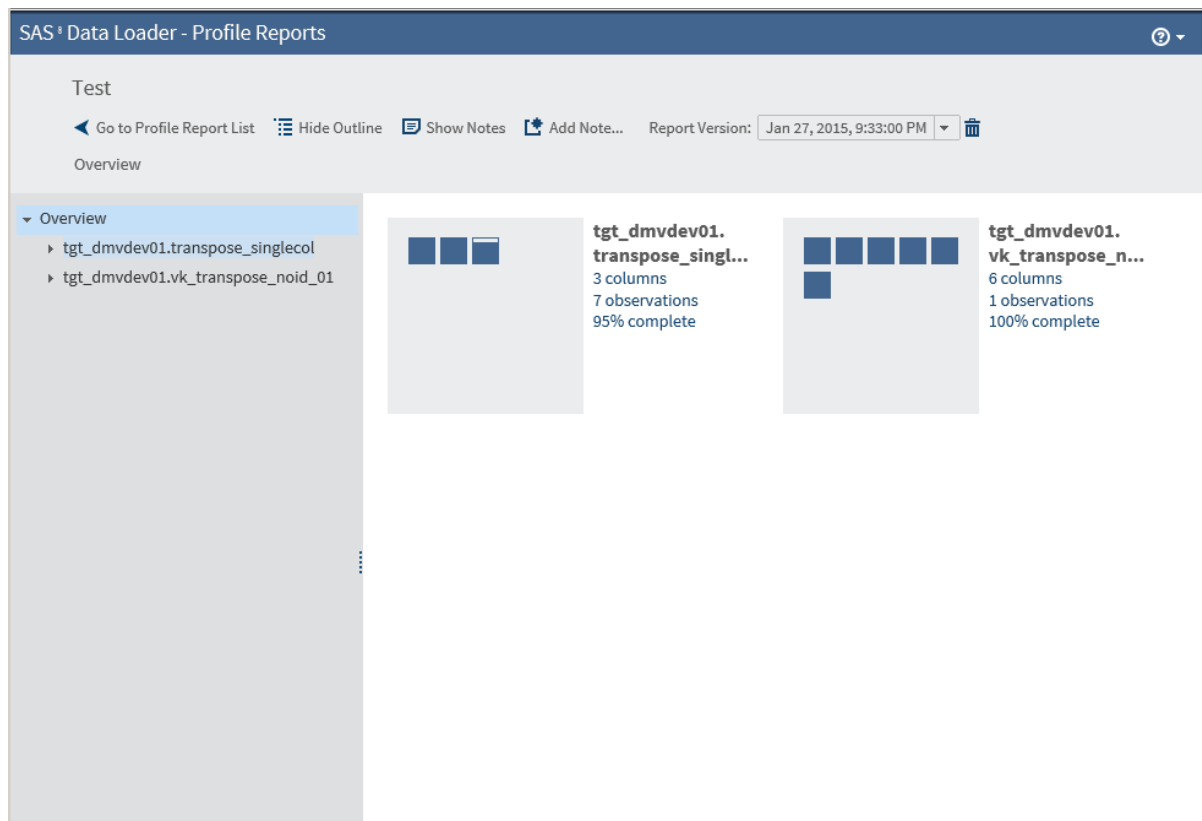
SAS® Data Loader - Profile Reports				
Select a Profile Report				
12 Profile Reports				
<div> <input type="text"/> </div>				
Name	Location	Last Run Date & Time Last Run Status		
<a href="#">abisen_customer_dimens...</a>	/data/sas/dmcontent_demo/Profiles/abisen_customer_dimensionVD...	1/2/2015, 11:48 AM	Succeeded	
<a href="#">Cust Bank Data Profile</a>	/data/sas/dmcontent_demo/Profiles/Cust_Bank_Data_ProfileVDPJO...	1/27/2015, 11:16 AM	Succeeded	
<a href="#">customer_adverse_accou...</a>	/data/sas/dmcontent_demo/Profiles/customer_adverse_accountsVD...	9/29/2014, 11:13 AM	Succeeded	
<a href="#">customer banking</a>	/data/sas/dmcontent_demo/Profiles/customer_bankingVDPJOB/cus...	1/1/2015, 4:11 PM	Succeeded	
<a href="#">customer dimension</a>	/data/sas/dmcontent_demo/Profiles/customer_dimensionVDPJOB/c...	11/28/2014, 9:25 AM	Succeeded	
<a href="#">prof</a>	/data/sas/dmcontent_demo/Profiles/profVDPJOB/profVDPJOB.xml	1/1/2015, 3:40 PM	Succeeded	
<a href="#">Profile Test - Client</a>	/data/sas/dmcontent_demo/Profiles/Profile Test - ClientVDPJOB/Pro...	1/23/2015, 12:47 PM	Succeeded	
<a href="#">Profile Test - Client Info</a>	/data/sas/dmcontent_demo/Profiles/Profile Test - Client InfoVDPJOB...	1/23/2015, 12:44 PM	Succeeded	
<a href="#">SASHELP_CLASS Profile</a>	/data/sas/dmcontent_demo/Profiles/SASHELP_CLASS_ProfileVDPJO...	12/4/2014, 8:19 AM	Succeeded	
<a href="#">smc_pro_test1</a>	/data/sas/dmcontent_demo/Profiles/smc_pro_test1VDPJOB/smc_pr...	12/10/2014, 11:08...	Succeeded	
<a href="#">Test</a>	/data/sas/dmcontent_demo/Profiles/TestVDPJOB/TestVDPJOB.xml	1/27/2015, 9:33 PM	Succeeded	
<a href="#">Test1</a>	/data/sas/dmcontent_demo/Profiles/Test1VDPJOB/Test1VDPJOB.xml	1/26/2015, 3:48 PM	Succeeded	

- 2 You can filter the list of reports using the following methods:

- Click  and select a date. This filter displays profile reports that were generated on or after the selected date.
  - Enter a text string into the search field.
  - Click  to remove the filter and restore the full list.
- 3 To delete profile reports, select one or more reports and click  .
- 4 To open a profile report, click its name.
- If the report contains a single table, the table opens directly in the detail view shown in [Step 6](#).
  - If the report contains multiple tables, the table opens in an overview:



- 5 You can click a table to go directly to a more detailed view or you can click  to open the outline view:



The following actions are available:

**Go to Profile Report List**

returns you to the Profile Report List.

**Show Outline**

displays or hides the outline in the left pane.

**Show Notes**

displays or hides notes in the right pane. You can filter the notes by entering a text string into the filter field.

**Add Note**

opens a dialog box in which you can add a note.

**Report Version**

enables you to select the version of the report by date.

- 6 Select a table in the **Overview** pane or click directly on the table icon to display detailed table information in the right pane. The Data Quality Metrics are displayed by default.

SAS Data Loader - Profile Reports

Test

Go to Profile Report List

Hide Outline

Show Notes

Add Note...

Report Version: Jan 27, 2015, 9:33:00 PM

Overview > tgt\_dmvdev01.transpose\_singlecol

Overview

tgt\_dmvdev01.transpose\_singlecol

tgt\_dmvdev01.vk\_transpose\_noid\_01

Count: 7

Data Quality Metrics

Column	#	Unique (n)	Unique (%)	Pattern (n)	Pattern (%)	Null (n)	Null (%)	Blank (n)	Blank (%)
tester1	1	7	100	2	28	0	0	0	0
_name_	2	1	14	1	14	0	0	0	0
col1	3	3	50	1	14	1	14	0	0


\* indicates data not available or not applicable for this column.

Descriptive Measures

Metadata Measures

Charts

**Note:** Currently, profile jobs count blank values, which consist of a series of blank-space characters, as SQL NULL values. In the profile report, the columns Blank (n) and Blank (%) are not populated at this time. The columns NULL (n) and NULL (%) reflect a summary of NULL and blank values in the profiled table.

- Click  next to a table name to display columns. Select a column to display detailed column information in the right pane:

SAS® Data Loader - Profile Reports

Test

Go to Profile Report List Hide Outline Show Notes Add Note... Report Version: Jan 27, 2015, 9:33:00 PM

Overview > tgt\_dmvdev01.transpose\_singlecol > col1

Overview

- tgt\_dmvdev01.transpose\_singlecol
  - \_name\_
  - col1
  - tester1
- tgt\_dmvdev01.vk\_transpose\_noid\_01

Count: 7

Standard Metrics

Unique (n)	3	Mean	(not applicable)
Unique (%)	50	Median	(not applicable)
Pattern (n)	1	S. D.	(not applicable)
Pattern (%)	14.29	S. E.	(not applicable)
Null (n)	1	Mode	25
Null (%)	14.3	Min. Value	20
Blank (n)	0	Max. Value	25
Blank (%)	0	Decimal Places	(not specified)

Ordinal Position 3

Data Type VARCHAR

Actual Type integer

Data Length 8192 chars

Nullable (not specified)

P.K. Candidate No


Min. Length 2

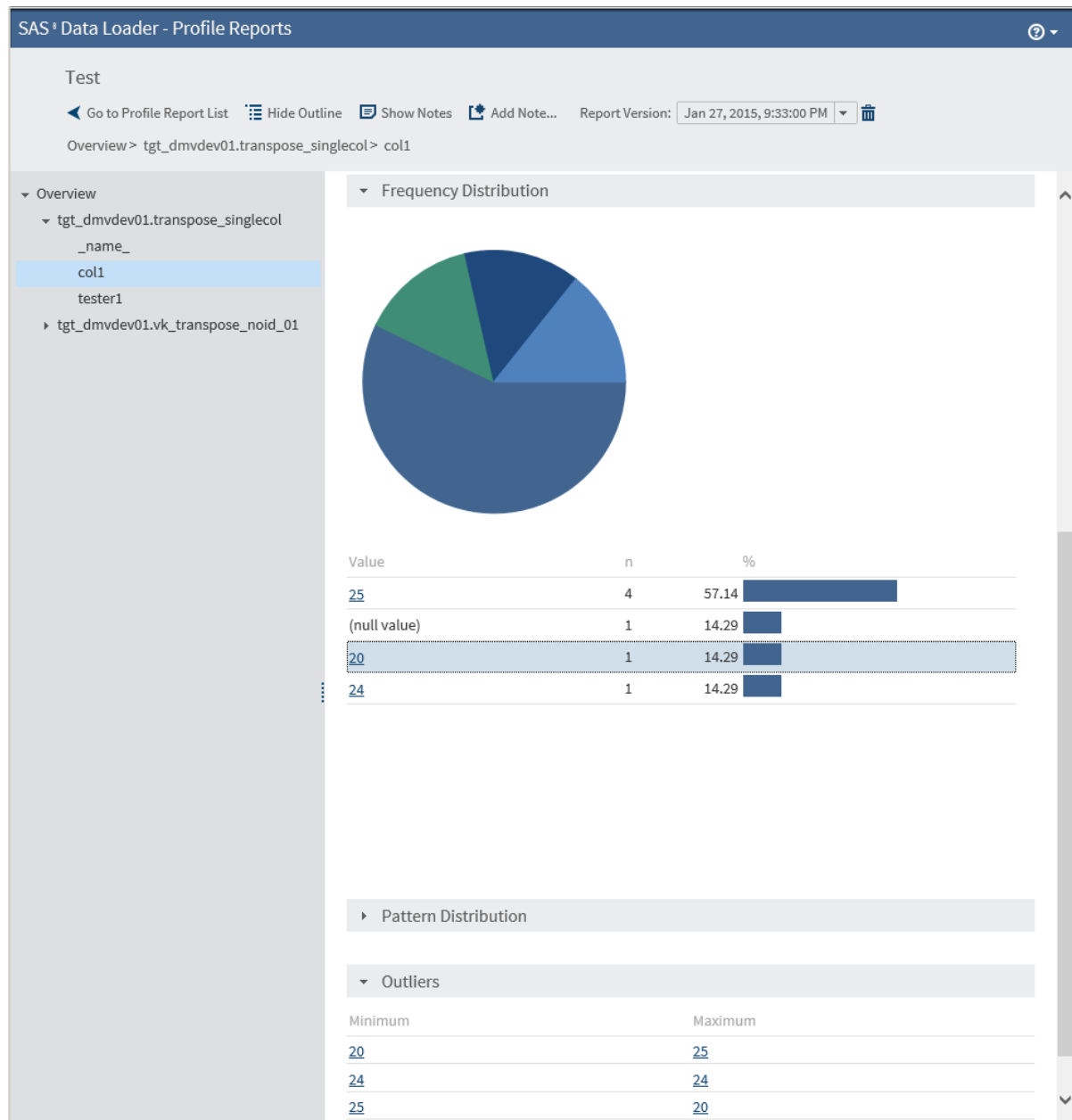
Max. Length 2

Frequency Distribution

Pattern Distribution

Outliers

- 8 Click  in the gray header bars to display the metrics in those sections. For example, clicking on Frequency Distribution icon displays the following metrics.



Clicking links in the detail view opens the [SAS Table Viewer](#).





## 5

## Copy Data To and From Hadoop

<b>Overview of the Copy Data Directives</b>	<b>69</b>
<b>Copy Data to Hadoop</b>	<b>70</b>
Introduction	70
Example	70
Install JDBC Drivers and Add Database Connections	81
Usage Notes	83
<b>Copy Data from Hadoop</b>	<b>84</b>
Introduction	84
Using Copy Data from Hadoop	84
About Drivers and Connections	93
Usage Notes	93
<b>Load Data to LASR</b>	<b>95</b>
Introduction	95
Prerequisites	95
Example	95
Connect to a SAS LASR Analytic Server Grid	96
Configure SSH Keys on a SAS LASR Analytic Server Grid	98
Usage Notes	98

---

### Overview of the Copy Data Directives

The Copy Data to Hadoop and Copy Data from Hadoop directives enable you to move data from your database management systems into and out of the Hadoop database.

The Copy Data to Hadoop and Copy Data from Hadoop directives use JDBC drivers to connect from your client machine to databases. The JDBC drivers on your client machine must be the same as those that are already on the Hadoop cluster. See [“Install JDBC Drivers and Add Database Connections”](#) on page 81 for more information.

---

## Copy Data to Hadoop

### Introduction



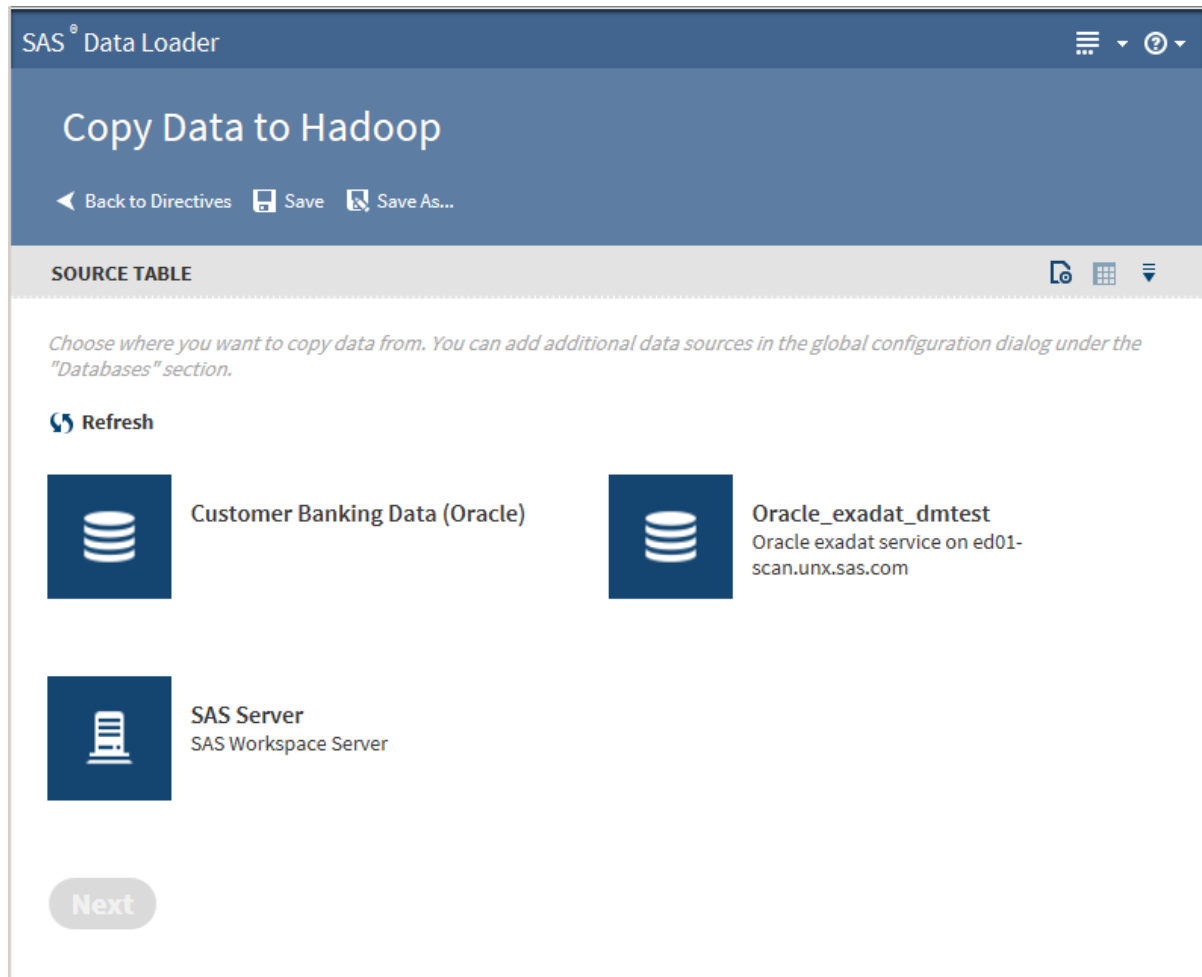
**Copy Data to Hadoop**  
Copy data from a database  
into Hadoop

The Copy Data to Hadoop directive enables you to copy data from your database management systems into Hadoop. You can also copy SAS data into Hadoop.

### Example

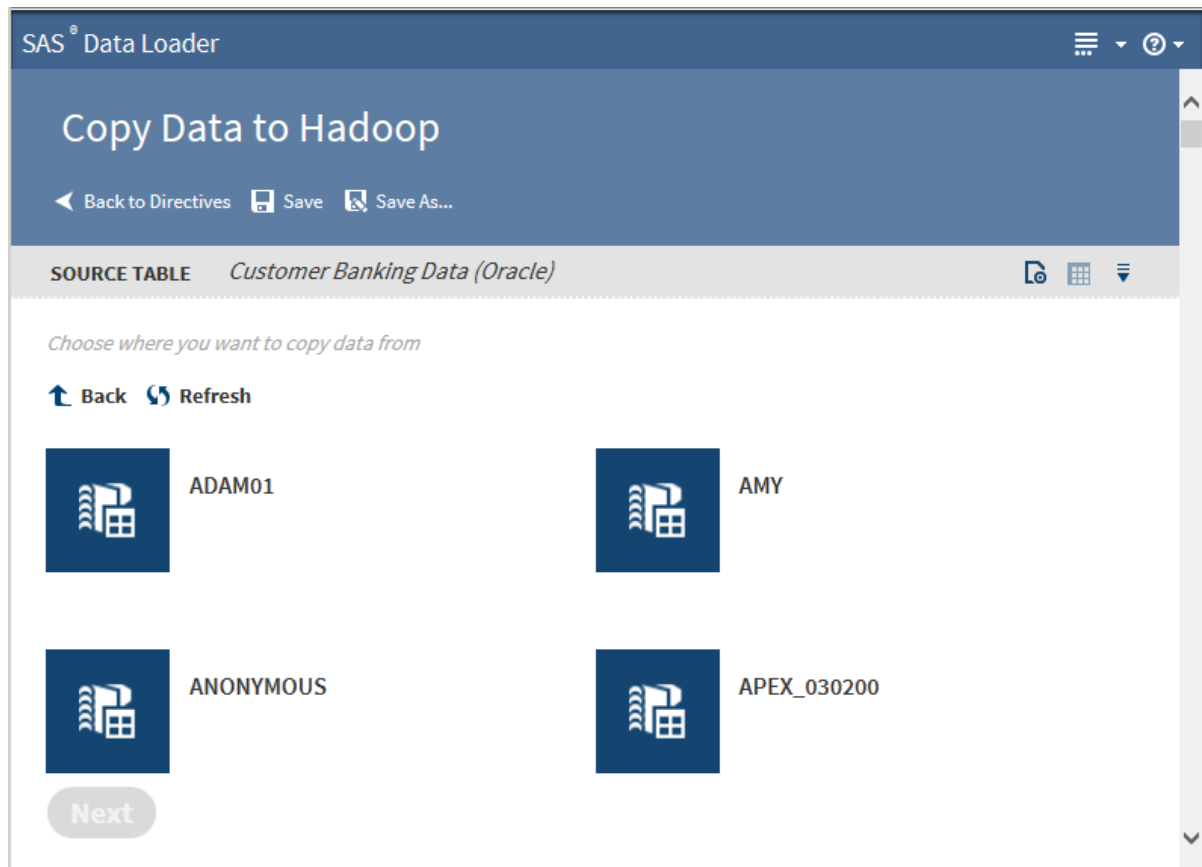
Follow these steps to copy data into Hadoop from a database:

- 1 On the SAS Data Loader page, click the Copy Data to Hadoop directive. The Source Table page that lists available databases is displayed:

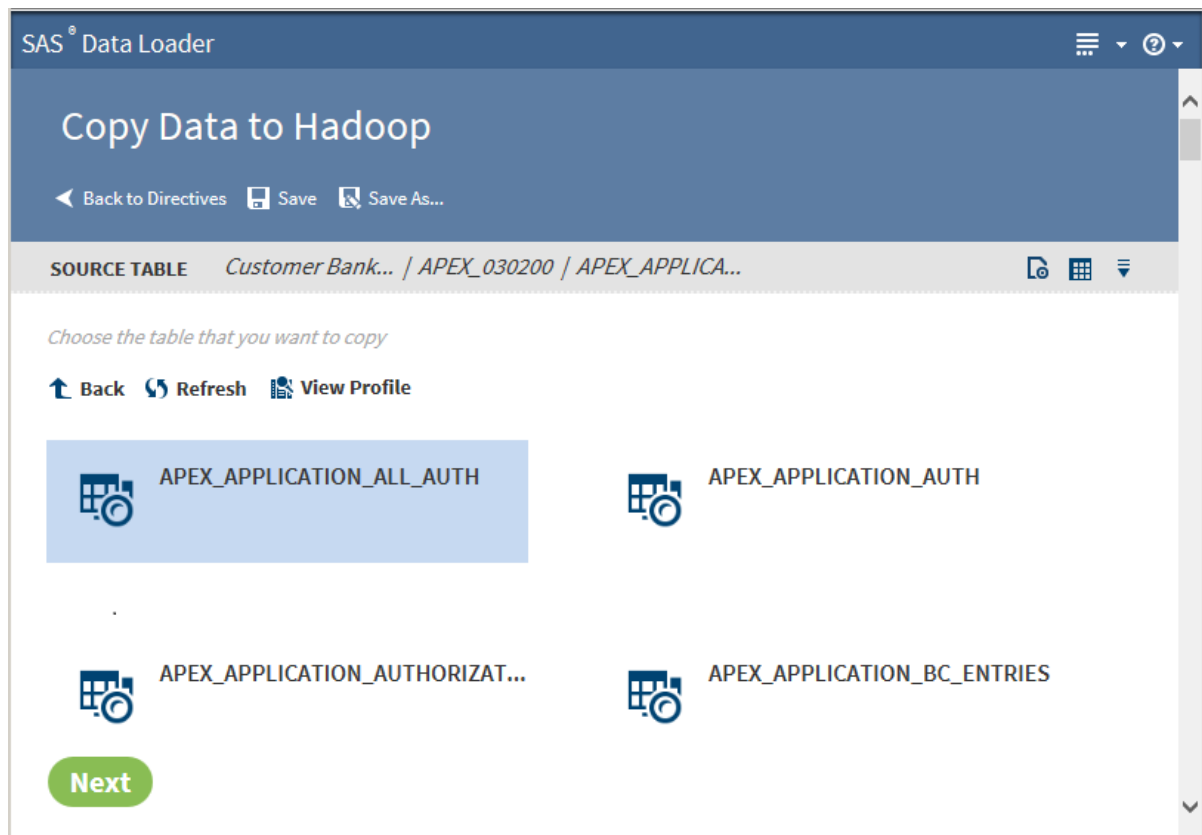


Note that the **SAS Server** data source points to the following location on the vApp host: `vApp-shared-folder/SASData/SAS Data Location`. To copy SAS data to Hadoop, all source tables must first be copied to this location.

- 2 Click a database to display its data sources:



3 Click a data source to display its tables:



#### 4 Select the table from which to copy data.


**TIP** If a profile already exists for a table, PROFILED appears beneath the table name. You can view the existing profile by selecting the table and clicking **View Profile**.

Clicking the Action menu  enables the following actions:

##### Open

opens the current directive page.

##### Table Viewer

enables you to view sample data from a table. Select the table, and then click  to display the [SAS Table Viewer](#).

##### Advanced Options

opens a dialog box that enables you to modify the advanced options. The advanced options enable additional character variable length to accommodate converted non-UTF8 encoding.

**TIP** It is recommended that you use UTF8 encoding in SAS data when copying data from SAS to Hadoop. The vApp always uses UTF8 encoding. If you copy a non-UTF8 encoded data set from elsewhere, then the Hadoop target table is not able to accommodate all the characters. This limitation is due to the increased number of bytes when the data is converted to UTF8 encoding.

**Note:** Modify only one of the following two advanced options. If you fill in both fields, then the value in the multiplier field is ignored.

Number of bytes to add to length of character variables (0 to 32766)  
Enter an integer value from 0 to 32766.

Multiplier to expand the length of character variables (1 to 5)  
Enter an integer value from 1 to 5.

Click **Next**. The Filter Rows page is displayed:

SAS Data Loader

## Copy Data to Hadoop

Back to Directives Save Save As...

SOURCE TABLE *Customer Ban... / APEX\_030200 / APEX\_APPLICA...*

FILTER ROWS *APPLICATION\_NAME = Testware*

Select the rows you want to filter.

☐ All rows ☒ Include rows where all of these rules apply:

Column: Operator: Value:

☒ Case sensitive

+ Add Rule

Next

- 5 The Filter Rows page enables you to filter the rows to be copied. You can select **All rows** or create filter rules. To create filter rules:
  - a Select **Include rows where all of these rules apply**.
  - b Select a column and an operator from the drop-down lists.
  - c Enter a value in the **Value** field.
  - d If appropriate, select **Case sensitive** for a string value.
  - e If you want to filter with additional rules, click **Add Rule**.

Click **Next**. The Columns page is displayed:

SAS Data Loader

## Copy Data to Hadoop

[Back to Directives](#)
[Save](#)
[Save As...](#)

**SOURCE TABLE** *Customer Ban... | APEX\_030200 | APEX\_APPLICA...*

**FILTER ROWS** *APPLICATION\_NAME = Testware*

**COLUMNS** *7 of 8 columns*

Select the columns you want to include in the target data file

☐ All columns
 ☒ Specify columns

Available columns:

- WORKSPACE
- APPLICATION\_ID
- APPLICATION\_NAME
- PAGE\_ID
- COMPONENT\_TYPE
- COMPONENT\_NAME
- AUTHORIZATION\_SCHEME
- STATUS

Selected columns:

- WORKSPACE
- APPLICATION\_ID
- APPLICATION\_NAME
- PAGE\_ID
- COMPONENT\_TYPE
- COMPONENT\_NAME
- AUTHORIZATION\_SCHEME

Next

- 6 The Columns page enables you to choose the columns to be copied. You can select **All columns** or **Specify columns**.

The columns in the **Selected columns** pane are copied to Hadoop. Select an individual column name and click or to move the column name between the **Available columns** pane and the **Selected columns** pane until the correct list of names appears in the **Selected columns** pane. Click or to move all column names at once.

When the column selection is complete, click **Next**. The Options page is displayed:

SAS® Data Loader

## Copy Data to Hadoop

◀ Back to Directives   Save   Save As...

<b>SOURCE TABLE</b>	Customer Bank... / APEX_030200 / APEX_APPLICA...	
<b>FILTER ROWS</b>	APPLICATION_NAME = Testware	
<b>COLUMNS</b>	7 of 8 columns	
<b>OPTIONS</b>	Processes: 1, Distribute Column: (Use default)	

*Specify how the copy operation will work. These defaults should only be changed for advanced scenarios. ?*

Number of processes:

Column used to distribute the copy:

**Next**

- 7 The values on the Options page should not be changed unless you have advanced knowledge of database operations.

**Note:** Changing the number of processes to greater than one expands the number of processes and source data connections that are used to import data. When running in this mode, a column must be identified in order to distribute the data across the parallel processes. This column is typically the primary key or index of the table in the data source. Only single columns are allowed. Numeric integer values that are evenly distributed in the data are recommended.

Click **Next**. The Target Table page is displayed with data sources:



The screenshot shows the SAS Data Loader interface for copying data to Hadoop. The title bar reads 'SAS® Data Loader'. The main heading is 'Copy Data to Hadoop'. Below the heading are navigation buttons: 'Back to Directives', 'Save', and 'Save As...'. The interface is divided into several sections:

- SOURCE TABLE**: *Customer Ban... / APEX\_030200 / APEX\_APPLIC...*
- FILTER ROWS**: *APPLICATION\_NAME = Testware*
- COLUMNS**: *7 of 8 columns*
- OPTIONS**: *Processes: 1, Distribute Column: (Use default)*
- TARGET TABLE**: This section is active and shows a list of target tables with a 'Refresh' button and a 'Next' button at the bottom.

The 'TARGET TABLE' section displays the following tables:

Table Name
abisen_demo
abisen_source
auto_dev_bsl
auto_dev_src
auto_dev_tgt
cloudera_manager_metastore...

8 Click a target data source to display its tables:

SAS® Data Loader

## Copy Data to Hadoop

◀ Back to Directives   Save   Save As...

**SOURCE TABLE** *Customer Ban... / APEX\_030200 / APEX\_APPLIC...*

**FILTER ROWS** *APPLICATION\_NAME = Testware*

**COLUMNS** *7 of 8 columns*

**OPTIONS** *Processes: 1, Distribute Column: (Use default)*

**TARGET TABLE** *auto\_dev\_tgt / rdc\_os\_info*

Select the target table you want to write the transformed data to.

↑ Return to data sources  
 ↻ Refresh  
 ✚ New Table...  
 👤 View Profile  
 Insert into the existing table ▼

**rdc\_os\_info**  
 NEW  
 Imported by sqoop on 2014/11/21 11:26:51

**rdc\_physical\_rdc**  
 Imported by sqoop on 2014/11/21 13:39:53

**rdc\_processor\_info**  
 Imported by sqoop on 2014/11/21 13:44:01

**test\_parse**

**Next**

- 9 Select the target table to which to copy data.

### TIP


- You can create a new table by clicking **New Table**
- If a profile already exists for a table, PROFILED appears next the table icon. You can view the existing profile by selecting the table and clicking **View Profile**.

Clicking the Action menu  enables the following actions:

### Open

opens the current directive page.

**Table Viewer**

enables you to view sample data from a table. Select the table, and then click  to display the [SAS Table Viewer](#). See

**Advanced Options**

opens a dialog box that enables you to modify the following advanced options:

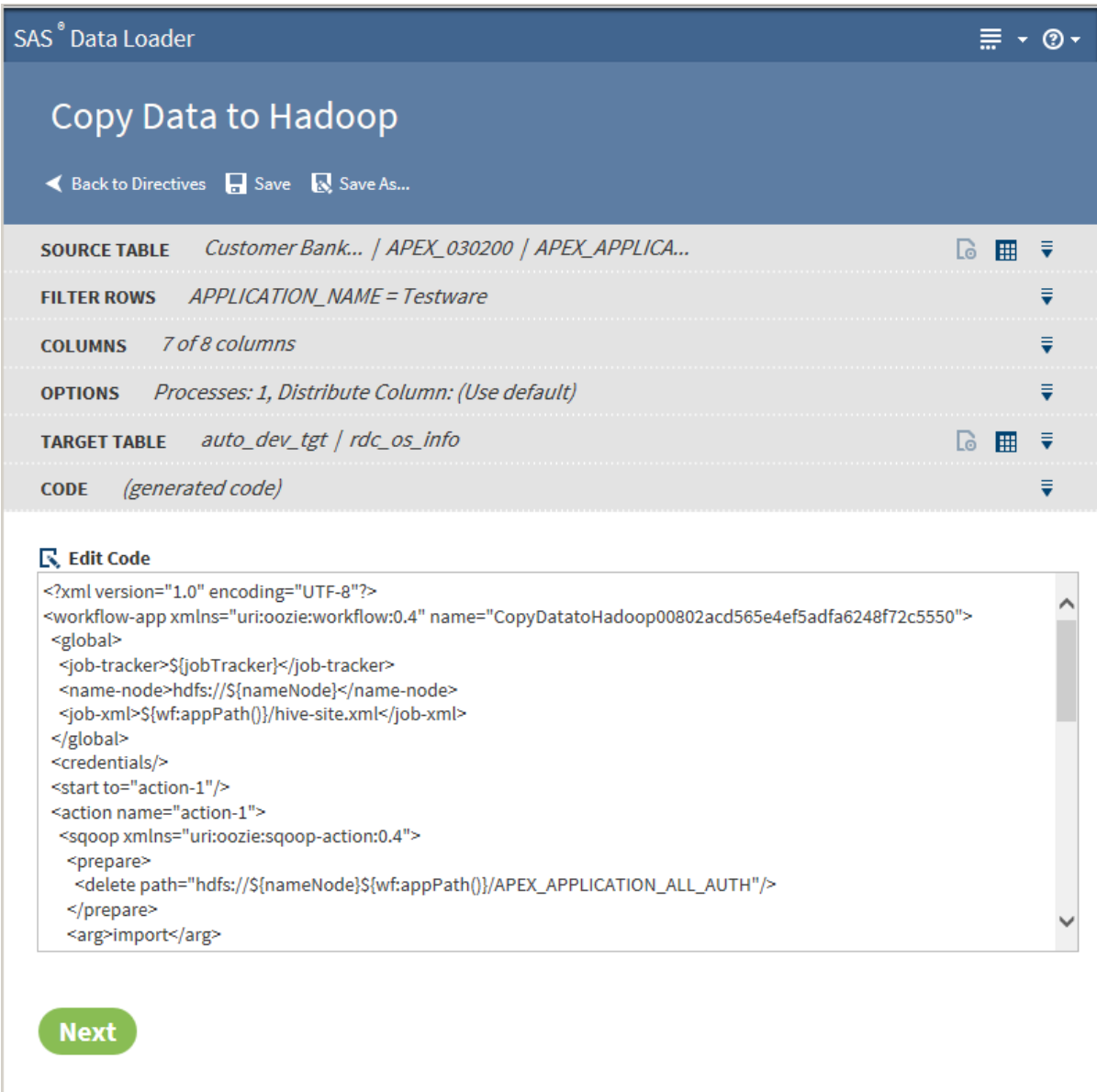
**Output table format**

Use the drop-down list to select one of five output table formats: Hive default, Text, Parquet, Orc, or Sequence.

**Delimiter**



Use the drop-down list to select one of five output table formats: Hive default, Comma, Tab, Space, or Other.





Click **Next**. The Code page is displayed:




SAS® Data Loader

## Copy Data to Hadoop

◀ Back to Directives  Save  Save As...

<b>SOURCE TABLE</b>	Customer Bank... / APEX_030200 / APEX_APPLICA...	  ▼
<b>FILTER ROWS</b>	APPLICATION_NAME = Testware	▼
<b>COLUMNS</b>	7 of 8 columns	▼
<b>OPTIONS</b>	Processes: 1, Distribute Column: (Use default)	▼
<b>TARGET TABLE</b>	auto_dev_tgt / rdc_os_info	  ▼
<b>CODE</b>	(generated code)	▼

 **Edit Code**

```
<?xml version="1.0" encoding="UTF-8"?>
<workflow-app xmlns="uri:oozie:workflow:0.4" name="CopyDatatoHadoop00802acd565e4ef5adfa6248f72c5550">
  <global>
    <job-tracker>${jobTracker}</job-tracker>
    <name-node>hdfs://${nameNode}</name-node>
    <job-xml>${wf:appPath()}/hive-site.xml</job-xml>
  </global>
  <credentials/>
  <start to="action-1"/>
  <action name="action-1">
    <sqoop xmlns="uri:oozie:sqoop-action:0.4">
      <prepare>
        <delete path="hdfs://${nameNode}${wf:appPath()}/APEX_APPLICATION_ALL_AUTH"/>
      </prepare>
      <arg>import</arg>
    </sqoop>
  </action>
</workflow-app>
```

**Next**

**10** Click **Edit Code** to modify the generated code. To cancel your modifications, click **Reset Code**.

**CAUTION!** Code edits are intended to be used only to support advanced features. Code edits are not needed or required under normal circumstances.

11 Click **Next**. The Result page is displayed:

SAS® Data Loader

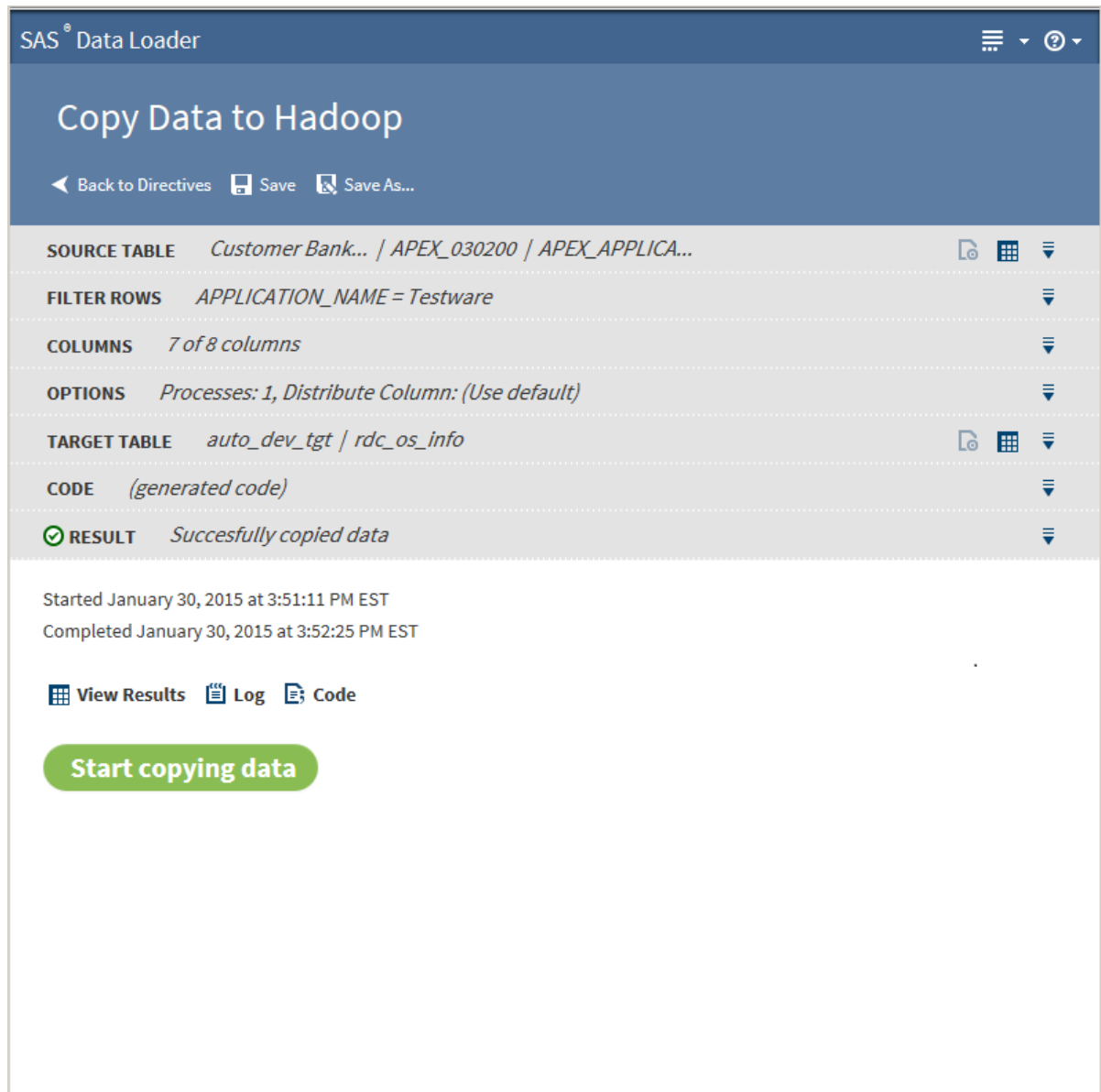
## Copy Data to Hadoop

◀ Back to Directives   Save   Save As...

SOURCE TABLE	Customer Bank... / APEX_030200 / APEX_APPLICA...	📄 📊 ⌵
FILTER ROWS	APPLICATION_NAME = Testware	⌵
COLUMNS	7 of 8 columns	⌵
OPTIONS	Processes: 1, Distribute Column: (Use default)	⌵
TARGET TABLE	auto_dev_tgt / rdc_os_info	📄 📊 ⌵
CODE	(generated code)	⌵
RESULT		⌵

**Start copying data**

12 Click **Start copying data**. The Result page displays the results of the copy process:



The following actions are available:

**View Results**

enables you to view the results of the copy process in the [SAS Table Viewer](#).

**Log**

displays the SAS log that is generated during the copy process.

**Code**

displays the SAS code that copies the data.

## Install JDBC Drivers and Add Database Connections

The directives Copy Data to Hadoop and Copy Data from Hadoop use JDBC drivers to connect your vApp host to databases such as Oracle. The JDBC drivers that are installed in the shared folder of the vApp must be the same as

those that are installed on the Hadoop cluster. The process of copying drivers to your vApp host was part of the initial vApp configuration process, as addressed in the *SAS Data Loader for Hadoop: vApp Deployment Guide*.

Follow these steps to add a new JDBC driver, and to add a database connection for that new driver.

- 1 To obtain a new JDBC driver, ask your Hadoop administrator to copy the driver on your Hadoop cluster into a ZIP file and mail you the ZIP file. This process is described in the *SAS Data Loader for Hadoop: Administrator's Guide*.

- 2 Unzip the ZIP file as follows:

- a Right-click and select **Open with WinZip** or **Expand All**.
- b If you are using WinZip, click **Unzip**.

- 3 In Windows Explorer, open the directory that is designated as the Shared Folder for your vApp. Here is a typical path to the Shared Folder:

C:\Program Files\SAS Data Loader\2.2\SASWorkspace\JDBCDrivers


To find the path to your Shared Folder, open the VMware Player Pro window and select **Player** ► **Manage** ► **Virtual Machine Settings**. In the Virtual Machine Settings window, click the **Options** tab, and then click **Shared Folders** (in the **Settings** list.) On the right side, the path to the Shared Folder is provided in the **Host Path** column.



- 4 Restart the vApp so that it can pick up the new JDBC driver.

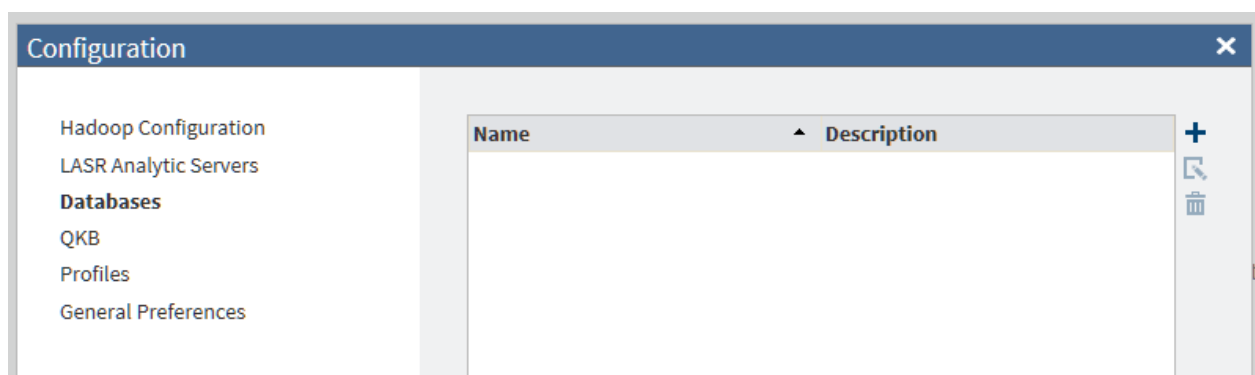
Check the “[Run Status](#)” directive to ensure that all jobs are stopped and saved.

In VMware Player Pro, select **Player** ► **Power** ► **Restart Guest**. Wait for the vApp to restart.

- 5 Open SAS Data Loader for Hadoop, as described in [Chapter 2, “Get Started,”](#) on page 5.

- 6 Click  and select **Configuration**.

- 7 In the Configuration window, click **Databases**. To add a new database connection, click **Add** . To edit an existing database connection, click the name of the connection, and then click **Edit** .



- 8 Contact your Hadoop administrator as needed to enter values into the Database Configuration window. The values of **Driver class** and **Connect string** are generated automatically when you select either Teradata or Oracle in the **Type** field. For an Oracle connection that requires a Service ID (SID), enter the SID in the **Database name** field. If you select **Other**, you must obtain these values from the JDBC driver provider.

- 9 When the configuration data is ready, click **Test Connection** to verify that the connection is operational.
- 10 If the test fails for a new Oracle connection, then examine the **Connect string** field. If the string has either of the following formats, then change the string to the other format and test the connection again.

```
jdbc:oracle:thin:@raintree.us.ourco.com:1521:oadev
```

```
jdbc:oracle:thin:@raintree.us.ourco.com:1521/oadev
```

One version uses a final colon character. The other version uses a final slash character.

To edit the **Connect string** field, click **Edit** .

- 11 Click **OK** to close the window.
- 12 Open the SAS Data Loader: Information Center and the SAS Data Loader for Hadoop and begin copying data to and from Hadoop with your new JDBC driver.

## Usage Notes

- See the usage note “[Changing the Default Maximum Length for SAS Character Columns](#)” on page 109.

- The `dfs.permissions.enabled` entry in the `hdfs-site.xml` file on the target Hadoop cluster must be set to `False`. If not, you might see an error message in the job history log. If you encounter an error, contact your Hadoop administrator.
- Error messages and log files that are produced by the Copy Data to Hadoop directive include the URL of the Oozie log file. Oozie is a job scheduling application that is used to execute Copy Data to Hadoop jobs. Refer to the Oozie log for additional troubleshooting information.
- If using Cloudera 5.2 or later with Teradata, the source Teradata table must have a primary key defined, or you must specify a column in **Column used to distribute the copy** on the [Options](#) page.
- SQL Server does not support the SQL standard syntax for specifying a Date literal, which is: `DATE 'date_literal'`. Edit the generated code and remove the word `DATE` that appears prior to the quoted date literal. For example, you would change `( table0.BEGDATE >= DATE '1990-01-01' )` to `( table0.BEGDATE >= '1990-01-01' )`. See [Step 10 on page 79](#) for information about the Code page.
- If using Hortonworks 2.1 or later with Teradata, creating a new table in Hadoop is not supported. You can insert only Teradata data into an existing table. See [Step 9 on page 78](#) for more information.
- When copying Oracle tables to Hadoop, Oracle table names must be uppercase.
- Data Loader directives will not create or replace hive tables when you are copying tables from Teradata into HortonWorks 2.1. This is due to a limitation in the HortonWorks Sqoop connector. One workaround is to ask your Hadoop administrator to drop any existing table, and then create an empty table with the desired schema. At that point, you could use the `APPEND` option in the Copy to Hadoop directive to copy a Teradata table into the empty table in Hortonworks 2.1.

---

## Copy Data from Hadoop

### Introduction



**Copy Data from Hadoop**  
Copy Data from Hadoop into  
a database

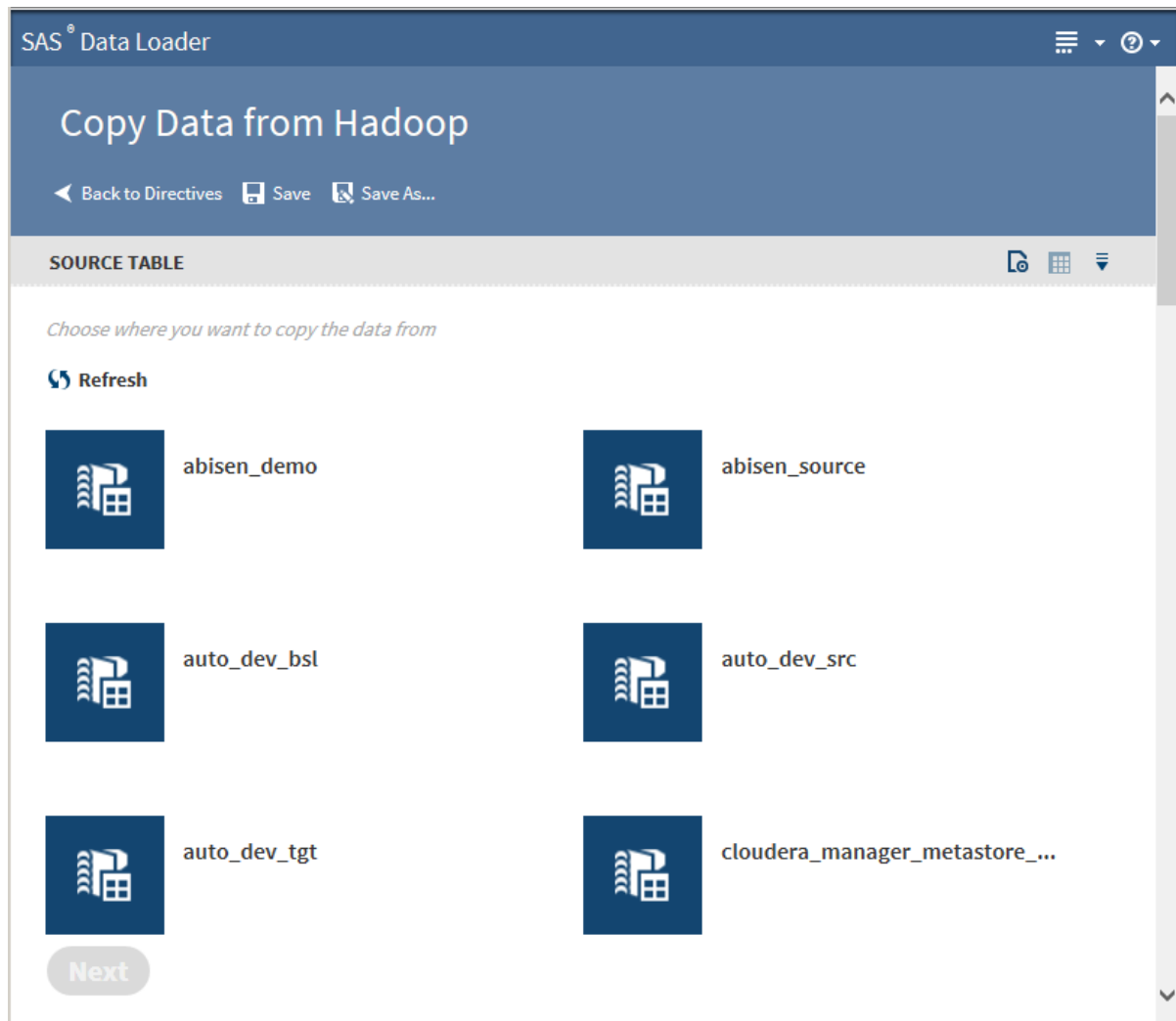
The Copy Data from Hadoop directive enables you to copy data from Hadoop into database management systems such as Oracle and Teradata.

### Using Copy Data from Hadoop

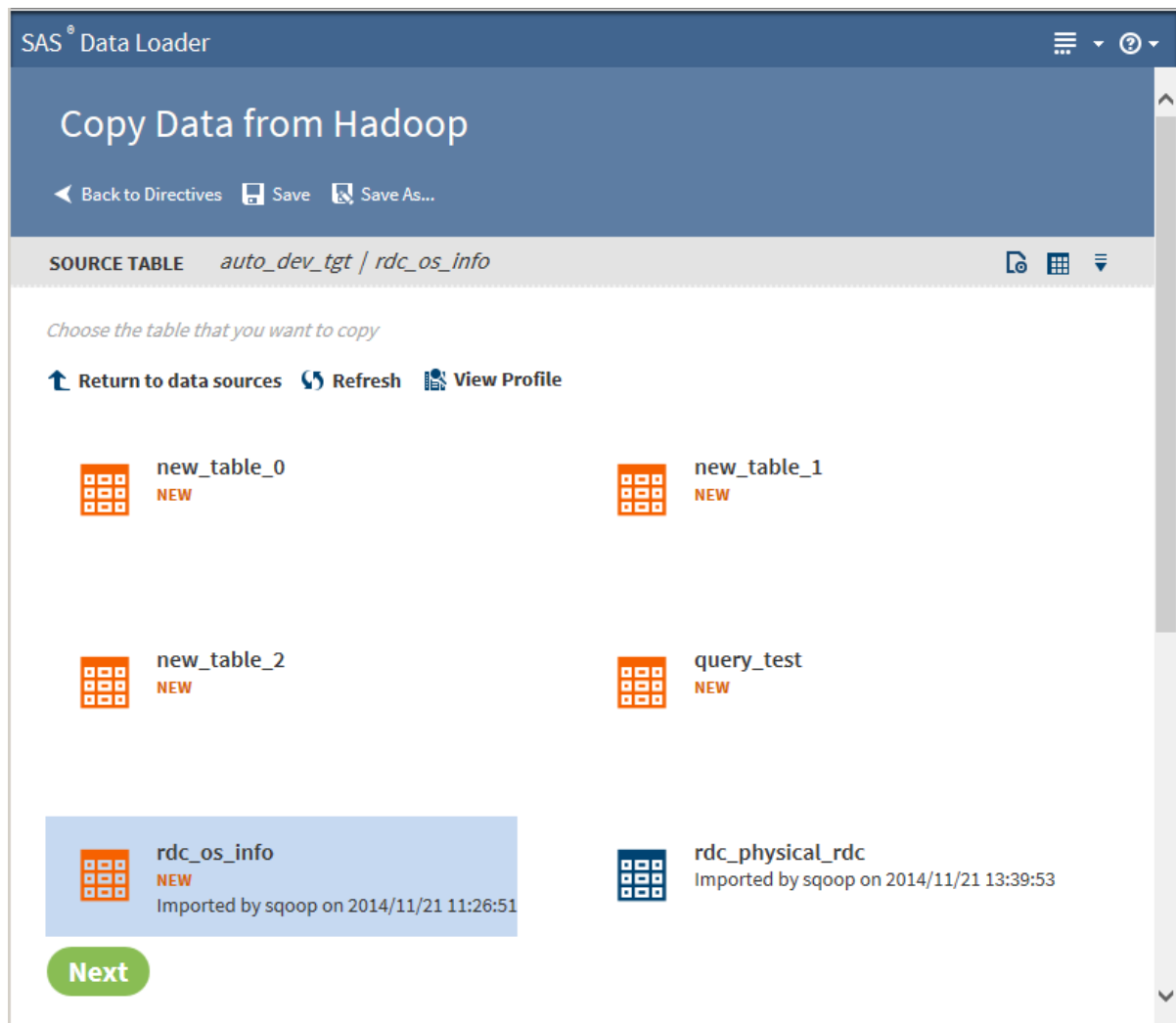
Follow these steps to copy data from Hadoop into a database:



- 1 On the SAS Data Loader page, click the Copy Data from Hadoop directive. The Source Table page that lists available data sources is displayed:



- 2 Click a data source to display its tables:




- 3 Select the table from which to copy data.

Clicking the Action menu  enables the following actions:

#### Open

opens the current directive page.

#### Table Viewer

enables you to view sample data from a table. Select the table, and then click  to display the [SAS Table Viewer](#).

#### Advanced Options

opens a dialog box that enables you to specify the maximum length for SAS columns. Entering a value here overrides the value specified in the **Configuration** options.

**Note:** If the source table has String data types, the resulting SAS data set could be very large. The length of the target field in the SAS data set is determined by the value of this option.

When table selection is complete, click **Next**. The Options page is displayed:

SAS® Data Loader

## Copy Data from Hadoop

◀ Back to Directives   Save   Save As...

**SOURCE TABLE**   *auto\_dev\_tgt / rdc\_os\_info*

**OPTIONS**   *Processes: 1*

*Specify how the copy operation will work. This default should only be changed for advanced scenarios.*

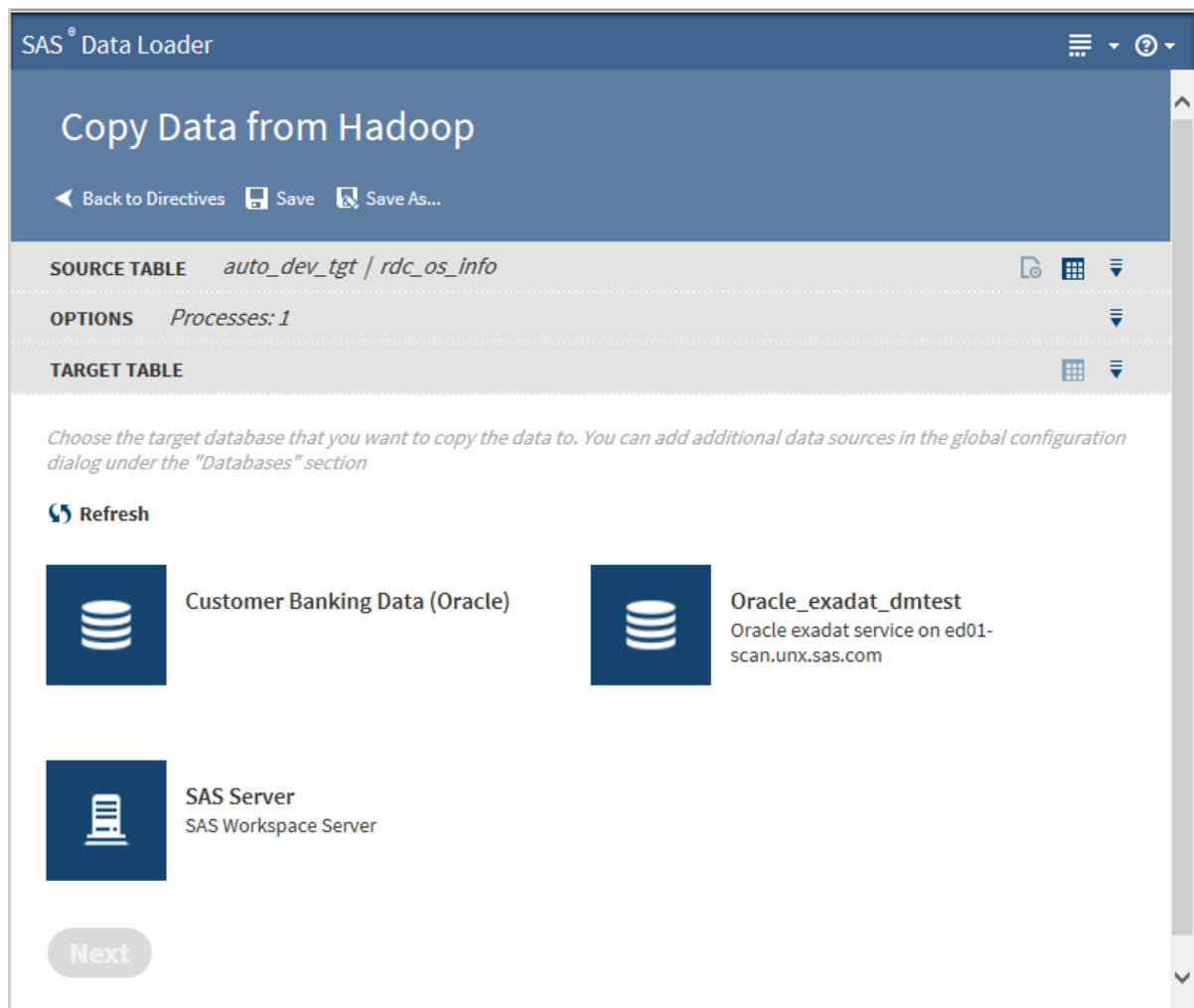
Number of processes:

**Next**

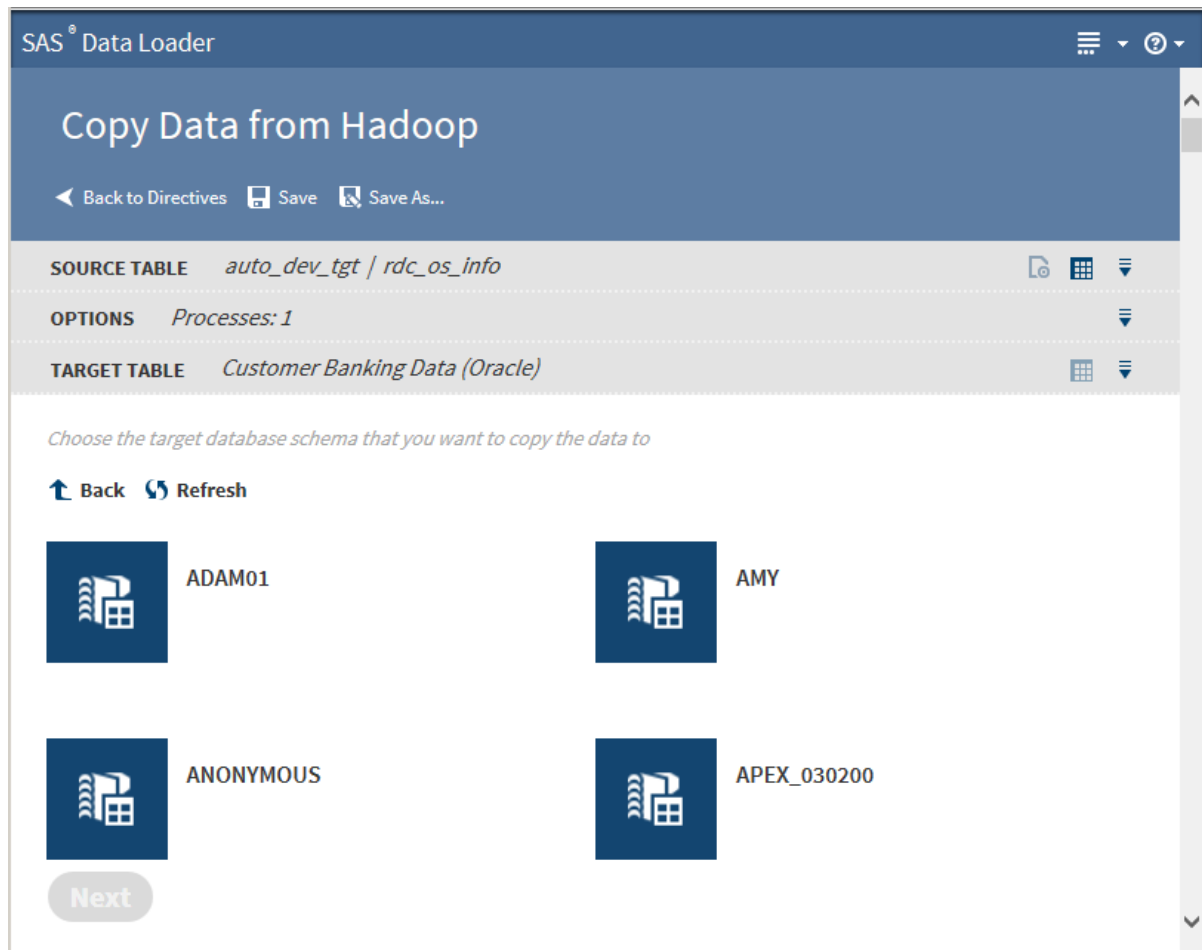
- 4 The value on the Options page should not be changed unless you have advanced knowledge of database operations.

**Note:** Changing the number of processes to greater than one expands the number of processes and source data connections that are used to import data.

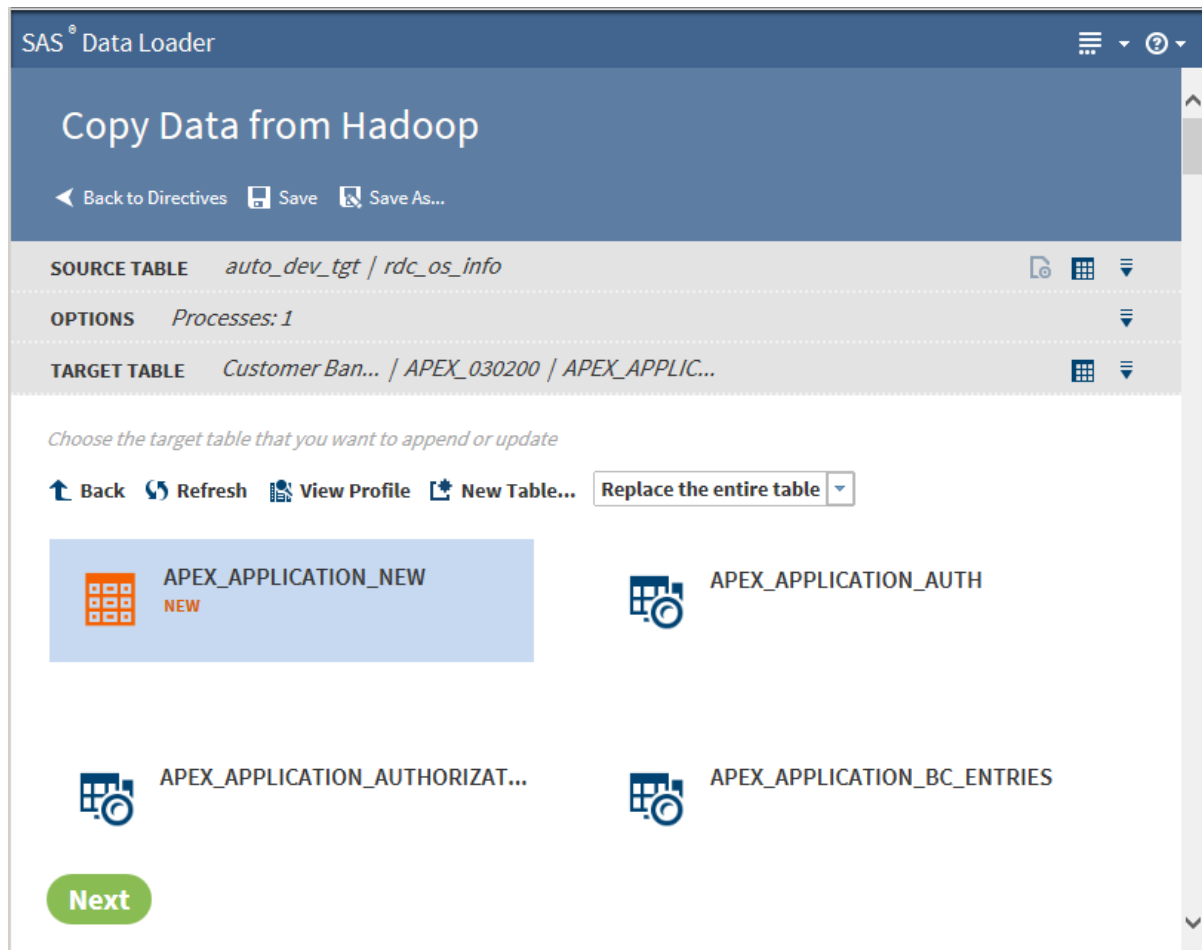
Click **Next**. The Target Table page is displayed with target databases:



5 Click a database to display its data sources:



6 Click a data source to display its tables:



- 7 Select the table from which to copy data.

#### TIP


- You can create a new table by clicking **New Table**
- If a profile already exists for a table, PROFILED appears next the table icon. You can view the existing profile by selecting the table and clicking **View Profile**.

Clicking the Action menu  enables the following actions:

#### Open

opens the current directive page.

#### Table Viewer

enables you to view sample data from a table. Select the table, and then click  to display the [SAS Table Viewer](#).

Click **Next**. The Code page is displayed:

SAS® Data Loader

## Copy Data from Hadoop

◀ Back to Directives   Save   Save As...

**SOURCE TABLE** *auto\_dev\_tgt / rdc\_os\_info*

**OPTIONS** *Processes: 1*

**TARGET TABLE** *Customer Ban... / APEX\_030200 / APEX\_APPLICA...*

**CODE** *(generated code)*

[Edit Code](#)

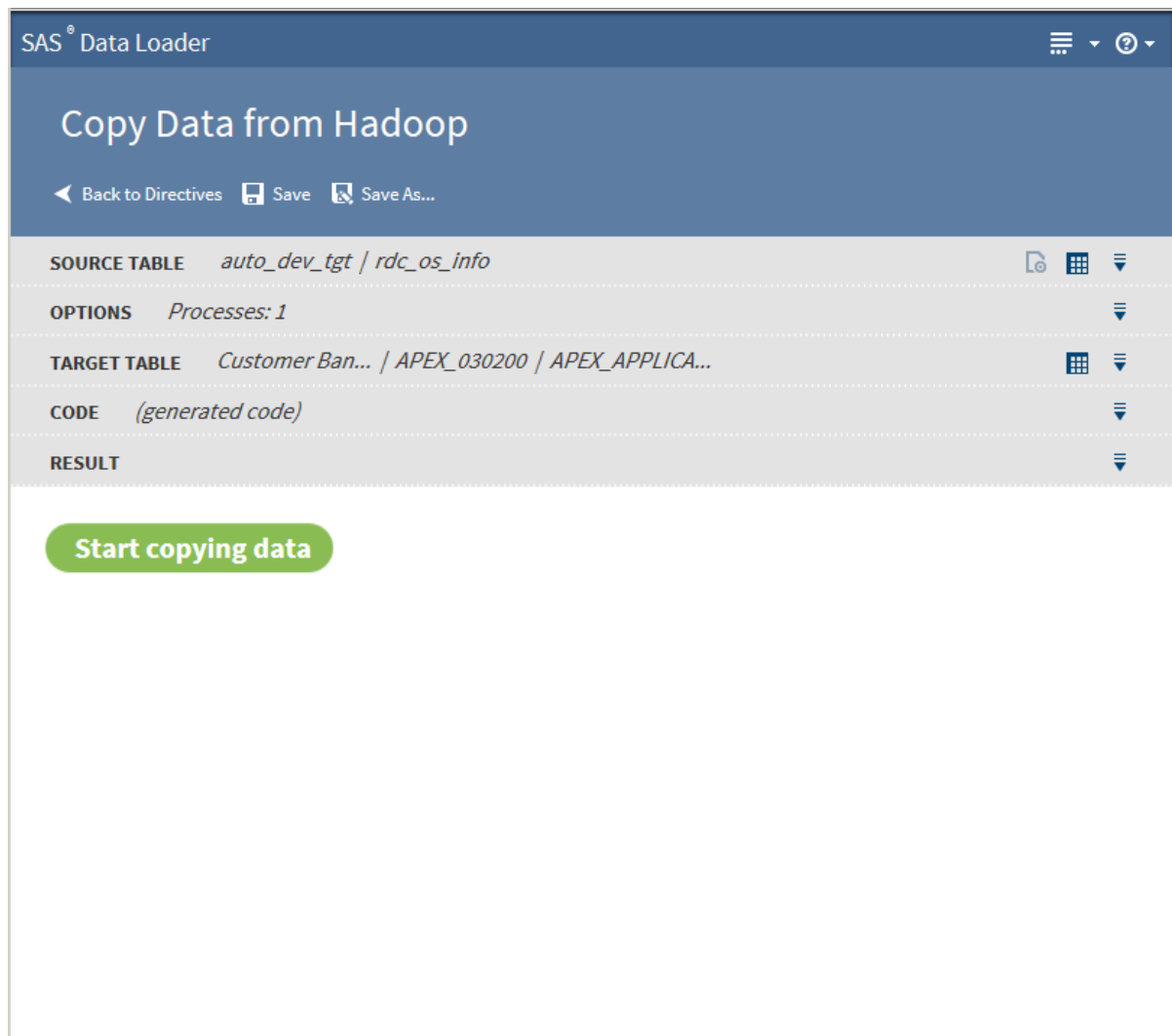
```
<?xml version="1.0" encoding="UTF-8"?>
<workflow-app xmlns="uri:oozie:workflow:0.4" name="CopyDatafromHadoop87f2449058d046cf8cfd0fefe056e5d3">
  <global>
    <job-tracker>${jobTracker}</job-tracker>
    <name-node>hdfs://${nameNode}</name-node>
    <job-xml>${wf:appPath()}/hive-site.xml</job-xml>
  </global>
  <credentials/>
  <start to="action-1"/>
  <action name="action-1">
    <sqoop xmlns="uri:oozie:sqoop-action:0.4">
      <arg>export</arg>
      <arg>--connect</arg>
      <arg>jdbc:oracle:thin:@flounder.na.sas.com:1521/ora112</arg>
      <arg>--username</arg>
```

**Next**

- 8 Click **Edit Code** to modify the generated code. To cancel your modifications, click **Reset Code**.

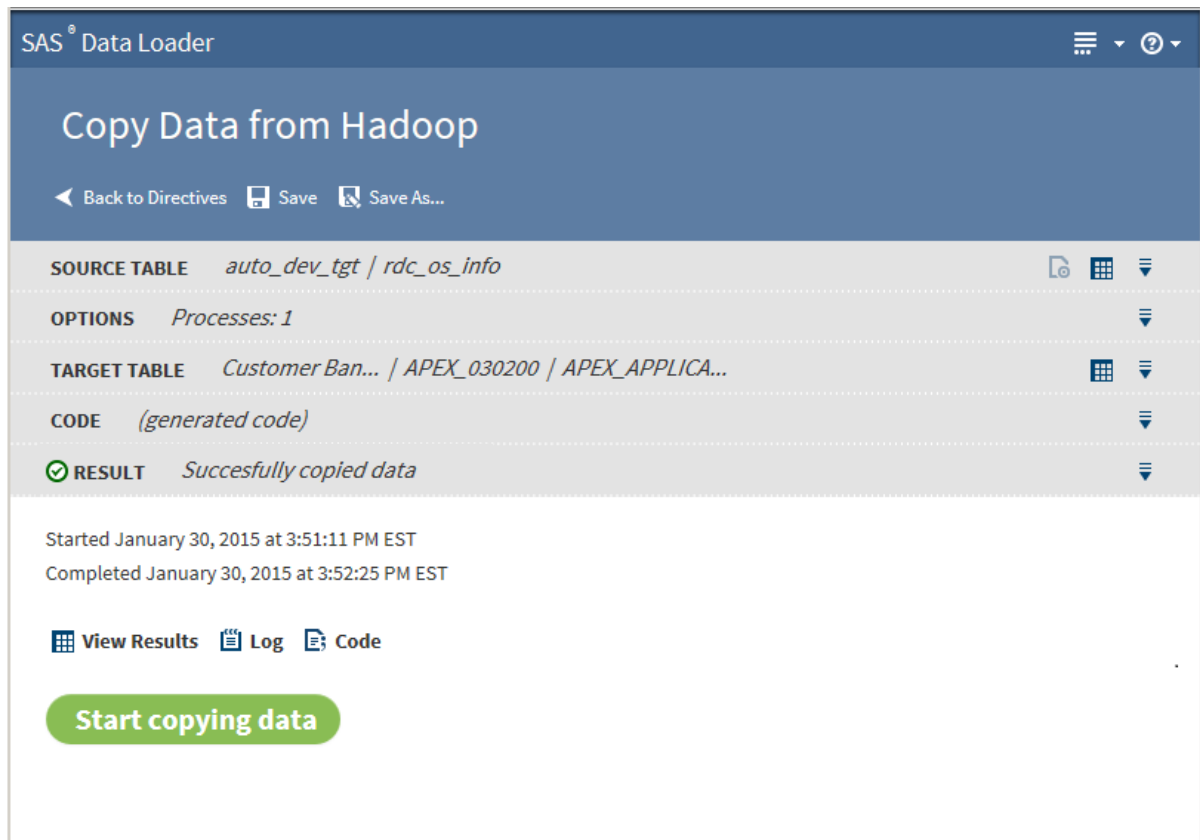
**CAUTION! Edit code only to implement advanced features.** Under normal circumstances, code edits are not need or required.

- 9 Click **Next**. The Result page is displayed:



- 10** Click **Start copying data**. The Result page displays the results of the copy process:





The following actions are available:

**View Results**

enables you to view the results of the copy process in the [SAS Table Viewer](#).the

**Log**

displays the SAS log that is generated during the copy process.

**Code**

displays the SAS code that copies the data.

## About Drivers and Connections

The Copy Data from Hadoop directive uses JDBC drivers to connect from your client machine to databases. The JDBC drivers on your client machine must be the same as those that are installed on the Hadoop cluster. See [“Install JDBC Drivers and Add Database Connections” on page 81](#) for more information.

## Usage Notes

- See the usage note [“Changing the Default Maximum Length for SAS Character Columns” on page 109](#).
- The hive.resultset.use.unique.column.names entry in the hive-site.xml file on the target Hadoop cluster must be set to False. If not, you might see an error message in the job history log. If you encounter an error, contact your Hadoop administrator.


- Source tables with a large number of columns can cause Copy From Hadoop jobs to fail. The job runs until the target table reaches the maximum number of columns that are supported in the target database. To resolve the problem, reduce the number of columns that are selected for the target and run the job again.
- If one or more varchar or string columns from a source Hadoop table contains more string data than the target database column, the Copy Data from Hadoop request times out. For example, a source Hadoop table might contain a string column named myString and a target Oracle table might contain a varchar(4000) column also named myString. If data in the Hadoop myString column has a length greater than 4000, then the copy request fails.
- When copying a Hadoop table to a database, a column name specified in the array of struct in the Hadoop table is not copied to the database table. This happens because of how structs are mapped to varchars in Sqoop.
- A copy from Hadoop is likely to fail if the name of a source column is also a reserved word in the target database.
- When copying a Hadoop table to Oracle, a mixed-case schema name generates an error.
- When copying a Hadoop table to Oracle, timestamp columns in Hadoop generate errors in Oracle. The Hive timestamp format differs from the Oracle timestamp format. To resolve this issue, change the column type in the Oracle target table from timestamp to varchar2.
- If a copy fails due to errors in null value formats, edit your code so that null string and null non-string arguments contain the value `null`:

```
<arg>--input-null-string</arg>
<arg>null</arg>
<arg>--input-null-non-string</arg>
<arg>null</arg>
<arg>--input-null-string</arg>
<arg>null</arg>
<arg>--input-null-non-string</arg>
<arg>null</arg>
```

- To copy Hadoop tables to Teradata, when the source contains a double-byte character set (DBCS) such as Chinese, follow these steps:

- 1 Edit the default connection string to include the option `charset=utf8`, as shown in this example:

```
jdbc:teradata://TeradataHost/Database=TeradataDB,charset=utf8
```

To edit the configuration string, open the Configuration window , click **Databases**, and click and edit the Teradata connection.

- 2 Ensure that the default character type for the Teradata user is UNICODE.
- 3 When a new Teradata table is created with the Copy Data from Hadoop directive, the column type for VARCHAR (and perhaps CHAR) columns should be set to CHARACTER SET UNICODE to accommodate wide characters.

---

# Load Data to LASR

## Introduction



### Load Data to LASR

Copy data from a source and load it into LASR. Existing data in the target table will be replaced

Use the Load Data to LASR directive to copy Hadoop tables to a grid of SAS LASR Analytic Servers. On the SAS LASR Analytic Servers, you can analyze tables using software such as SAS Visual Analytics.

**Note:** The Load Data to LASR directive is distinct and separate from the Load to LASR capability that is provided by the SAS LASR Analytic Server.

## Prerequisites

In order to use this directive, you must connect to a grid of SAS LASR Analytic Servers. Ask your SAS LASR Analytic Server administrator to verify that the following prerequisites have been met:




- A grid of SAS LASR Analytic Servers, release 2.5 or later, must be licensed, installed, and configured.
- SAS Visual Analytics 6.4 or later must be installed and configured on the SAS LASR Analytic Servers.
- The SAS LASR Analytic Servers must be registered on a SAS Metadata Server.
- The SAS LASR Analytic Servers must be configured to start automatically.
- The SAS LASR Analytic Servers must have memory and disk allocations that are large enough to accept Hadoop tables. The Load Data to LASR directive does not check the SAS LASR Analytic Servers for available memory or disk space.

After verifying the prerequisites above, ask your Hadoop administrator if your Hadoop cluster is secured with Kerberos. If so, you are ready to specify a connection to the SAS LASR Analytic Server grid. Follow the steps that are described in [“Connect to a SAS LASR Analytic Server Grid” on page 96](#).

If your Hadoop cluster is not secured with Kerberos, ask the SAS LASR Analytic Server administrator to configure Secure Shell (SSH) keys for SAS Data Loader on your SAS LASR Analytic Server grid. Direct your server administrator to the steps that are described in [“Configure SSH Keys on a SAS LASR Analytic Server Grid” on page 98](#).



## Example



Follow these steps to create and run the Load Data to LASR directive:

- 1 Open SAS Data Loader for Hadoop, as described in [Chapter 2, “Get Started,” on page 5](#).
- 2 In the Directives page, click **Load Data to LASR**.
- 3 In the Source Table page, click the schema that contains the source table that you want to load. Clicking the schema displays the tables in that schema. Click the table that you want to load onto your grid of SAS LASR Analytic Servers, and then click **Next**.
- 4 In the Target Table page, click the SAS LASR Analytic Server that you want to receive the target table. Clicking displays target table configuration fields and controls.
- 5 As needed, change the name in the **Target table name** field. The field defines the name of the table on the grid of SAS LASR Analytic Servers.
- 6 Select options as needed to replace any existing table of the same name or to compress the target table on the grid of SAS LASR Analytic Servers.
- 7 Click the **Locations** link to view or change the default storage options for the target table on the grid of SAS LASR Analytic Servers.
- 8 In the Locations window, you can change the SAS folder, the library name, and the required tag that accompanies the table name.
- 9 In the Target Table page, click **Next**.
- 10 In the Result page, click **Start loading data**. SAS proceeds to generate code for the directive and displays the **Code** icon . Click the icon to open or save the text of the SAS code that comprises the directive.
- 11 During the execution of the directive, the Result page displays the **Log** icon . Click the icon to open or save the SAS log file that is generated during the execution of the directive.
- 12 At the conclusion of the directive, the Result banner receives a status icon that indicates the success or failure of the directive. To view the target table on the SAS LASR Analytic Server, click the **View Results** icon .

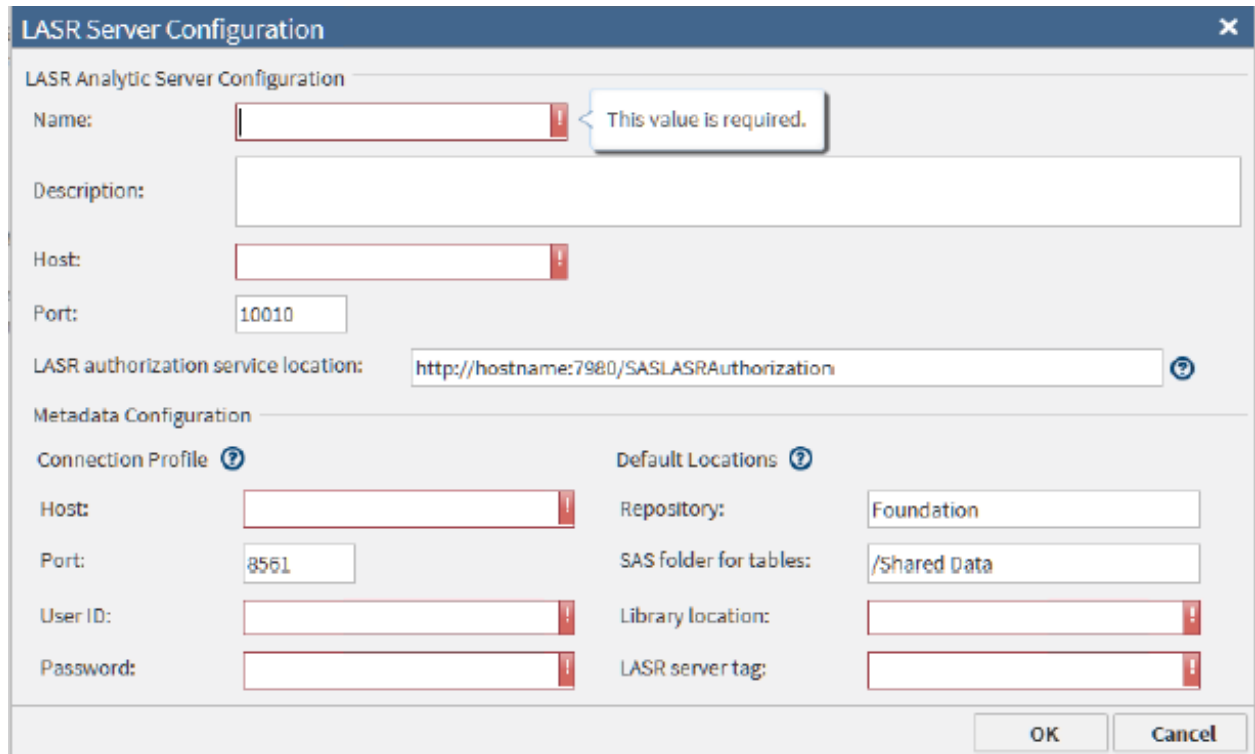
## Connect to a SAS LASR Analytic Server Grid

Your SAS LASR Analytic Server administrator can provide the information that you need to configure connections to a SAS LASR Analytic Server grid. Follow these steps to configure a connection:

- 1 Open SAS Data Loader for Hadoop, as described in [Chapter 2, “Get Started,” on page 5](#).
- 2 In the SAS Data Loader panel, click the More icon  and select **Configuration**.
- 3 Click **SAS LASR Analytic Servers**.
- 4 To configure a new SAS LASR Analytic Server, click the Add icon . If you are changing an existing server connection, click that connection in the list,

and then click the Edit icon . To delete a server connection, select it and click the Delete icon .

- 5 In the LASR Server Configuration window, enter or change your choice of server name and description in the **Name** and **Description** fields.



- 6 In the **Host** field, add or change the full network name of the host of the SAS LASR Analytic Server. A typical name is similar to lasr03.us.ourco.com.
- 7 In the **Port** field, add or change the number of the port that the SAS LASR Analytic Server uses to listen to connections from SAS Data Loader. The default port number is 10010.
- 8 In the field **LASR authorization service location**, add or change the HTTP address of the authorization service.
- 9 Under Connection Profile, in the lower of the two **Host** fields, add or change the network name of the SAS Metadata Server that is accessed by the SAS LASR Analytic Server.
- 10 In the lower of the two **Port** fields, add or change the number of the port that the SAS Metadata Server uses to listen for client connections. The default value 8561 is frequently left unchanged.
- 11 In the **User ID** and **Password** fields, add or change the credentials that SAS Data Loader will use to connect to the SAS Metadata Server. These values are stored in encrypted form.
- 12 In the **Repository** field, specify the name of the repository on the SAS LASR Analytic Server that will receive the downloads from Hadoop. The default value **Foundation** might suffice.

- 13 In the field **SAS folder for tables**, specify the path inside the repository that will contain the downloads from Hadoop. The default value `/sharedData` might suffice.
- 14 In the **Library location** field, add or change the name of the SAS library that will be referenced by the Load Data to LASR directive.
- 15 In the **LASR server tag** field, add or change the name of the tag that will be associated with each table that is downloaded from Hadoop. The tag is required. It is used along with the table name to uniquely identify tables that are downloaded from Hadoop.
- 16 Review your entries and click **OK** to return to the Configuration window.  
At this point, you can define or edit a connection to another SAS LASR Analytic Server.

## Configure SSH Keys on a SAS LASR Analytic Server Grid

If your Hadoop cluster is not secured with Kerberos, ask your SAS LASR Analytic Server administrator to configure Secure Shell (SSH) keys for SAS Data Loader on your SAS LASR Analytic Server grid. After that, you can configure a connection to the SAS LASR Analytic Server grid as described above.

The server administrator will perform these steps:

- 1 On the SAS LASR Analytic Server grid, the administrator must create the user `sasdlr1`, as described in the *SAS LASR Analytic Server: Reference Guide*.
- 2 The administrator must generate a public key and a private key for `sasdlr1` and install those keys, as described in the *SAS LASR Analytic Server: Reference Guide*.
- 3 The administrator must copy the public key file from SAS Data Loader at `vApp-install-path\vApp-instance\shared-folder \Configuration\sasdemo.pub`. A typical path is `C:\Program Files\SASDataLoader\dataloader-3p.22on94.1-devel-vmware.vmware(1)\dataloader-3p.22on94.1-devel-vmware\SASWorkspace\Configuration`.

Append the SAS Data Loader public key to the file `~sasdlr1/.ssh/authorized_keys` on the head node of the grid.

**CAUTION!** To maintain access to the SAS LASR Analytic Servers, you must repeat step 3 each time you replace your installation of SAS Data Loader for Hadoop.

It is not necessary to repeat this step if you update your vApp by clicking the **Update** button in the SAS Data Loader Information Center.

## Usage Notes

The Load Data to LASR directive moves entire tables. To improve performance, you can filter the rows and manage the columns before you load the table to the

SAS LASR Analytic Server grid. To reduce table size, use the directives Transform Data in Hadoop or Query Data in Hadoop.

See the usage note “ [Changing the Default Maximum Length for SAS Character Columns](#)” on page 109.





6

# Manage Jobs

- Overview of Job Management Directives* ..... 101
- Run Status* ..... 102
  - Introduction ..... 102
  - Using Run Status ..... 102
  - About Unsaved Jobs ..... 104
  - About Incomplete Jobs ..... 104
- Saved Directives* ..... 104
  - Introduction ..... 104
  - Opening Saved Directives ..... 104
  - Managing Saved Directories ..... 105

---

## Overview of Job Management Directives

The job management directives enable you to view the status of current and previous jobs and to modify and execute saved directories. The Run Status directive displays information about the current execution state of jobs. The Saved Directives directive enables you to open, edit, and manage your existing directives.

Here's an example of the Run Status directive:

SAS® Data Loader

## Run Status

◀ Back to Directives

Show: Last 30 Days Refresh Clear All

Name	Status	Start Time	End Time	Run Time
Transpose Data in Hadoop	Stopped	Jan 29, 2015, 10:50:41 AM	Jan 29, 2015, 10:58:59 AM	00:08:17.849
Profile Data	Successful	Jan 28, 2015, 10:46:16 AM	Jan 28, 2015, 10:51:05 AM	00:04:48.888
Profile Data	Successful	Jan 27, 2015, 9:33:16 PM	Jan 27, 2015, 9:38:06 PM	00:04:50.195
Transpose Data in Hadoop	Successful	Jan 27, 2015, 6:11:33 PM	Jan 27, 2015, 6:18:53 PM	00:07:19.842
Sort and De-Duplicate Data	Successful	Jan 27, 2015, 6:07:14 PM	Jan 27, 2015, 6:07:57 PM	00:00:43.532
Cleanse Data in Hadoop	Successful	Jan 27, 2015, 5:59:16 PM	Jan 27, 2015, 6:01:34 PM	00:02:18.229
Copy Data to Hadoop	Successful	Jan 27, 2015, 5:57:05 PM	Jan 27, 2015, 5:57:21 PM	00:00:15.634
Cleanse Data in Hadoop	Successful	Jan 27, 2015, 5:48:45 PM	Jan 27, 2015, 5:50:43 PM	00:01:58.435
Cleanse Data in Hadoop	Successful	Jan 27, 2015, 5:29:11 PM	Jan 27, 2015, 5:38:37 PM	00:09:25.453
Cleanse Data in Hadoop	Successful	Jan 27, 2015, 5:09:16 PM	Jan 27, 2015, 5:36:27 PM	00:27:10.629
Copy Data to Hadoop	Successful	Jan 27, 2015, 5:05:52 PM	Jan 27, 2015, 5:06:08 PM	00:00:16.021
Cleanse Data in Hadoop	Successful	Jan 27, 2015, 4:26:08 PM	Jan 27, 2015, 4:30:22 PM	00:04:14.029

## Run Status

### Introduction



#### Run Status

Show the status of current and previous directive executions

Use the Run Status directive to view job runs. Each run is listed with its current execution status, start time, end time, and run time. The Status column value can be In Progress, Stopped, Failed, or Successful.

### Using Run Status


In the SAS Data Loader page, click the Run Status directive. The Run Status page is displayed:

Name	Status	Start Time	End Time	Run Time
Transpose Data in Hadoop	Stopped	Jan 29, 2015, 10:50:41 AM	Jan 29, 2015, 10:58:59 AM	00:08:17.849
Profile Data	Successful	Jan 28, 2015, 10:46:16 AM	Jan 28, 2015, 10:51:05 AM	00:04:49.123
Profile Data	Successful	Jan 27, 2015, 9:33:16 PM	Jan 27, 2015, 9:38:06 PM	00:04:50.123
Transpose Data in Hadoop	Successful	Jan 27, 2015, 6:11:33 PM	Jan 27, 2015, 6:18:53 PM	00:07:20.123
Sort and De-Duplicate Data	Successful	Jan 27, 2015, 6:07:14 PM	Jan 27, 2015, 6:07:57 PM	00:00:43.123
Cleanse Data in Hadoop	Successful	Jan 27, 2015, 5:59:16 PM	Jan 27, 2015, 6:01:34 PM	00:02:18.123
Copy Data to Hadoop	Successful	Jan 27, 2015, 5:57:05 PM	Jan 27, 2015, 5:57:21 PM	00:00:16.123
Cleanse Data in Hadoop	Successful	Jan 27, 2015, 5:48:45 PM	Jan 27, 2015, 5:50:43 PM	00:01:58.435
Cleanse Data in Hadoop	Successful	Jan 27, 2015, 5:29:11 PM	Jan 27, 2015, 5:38:37 PM	00:09:25.453
Cleanse Data in Hadoop	Successful	Jan 27, 2015, 5:09:16 PM	Jan 27, 2015, 5:36:27 PM	00:27:10.629
Copy Data to Hadoop	Successful	Jan 27, 2015, 5:05:52 PM	Jan 27, 2015, 5:06:08 PM	00:00:16.021
Cleanse Data in Hadoop	Successful	Jan 27, 2015, 4:26:08 PM	Jan 27, 2015, 4:30:22 PM	00:04:14.029

By default, the Run Status page displays all of the directives that have run in the past 30 days. The most recent runs appear at the top of the list. You can change the default of 30 days by selecting a new value from the **Show** drop-down list. Reports are identified by the given name or by the generic name of the directive (for example, Transform Data in Hadoop.) Given names are created when you save a directive.

When you click **Refresh**, you receive updates for all running jobs, including any that were started or completed after you opened the Run Status page.

Clicking **Clear All** clears all of the reports from the Run Status page. Clearing reports permanently removes the reports from the vApp database.

Clicking the Action menu  for a job in the list enables the following actions:

#### Open

opens the directive associated with the job.

#### View Profile Report

for successful Profile Data jobs, enables you to view the Profile Report unless the report has been deleted from the Saved Profiles directive. See [“Saved Profile Reports” on page 62](#) for more information about the profile report.

#### View Results

for completed transformations or queries, enables you to view a sample of the target table in the [SAS Table Viewer](#).

#### Log

displays the SAS log that is generated during the execution of the profile job..

**Code**

displays the SAS code that is generated during the execution of the profile job..

**Start**

starts a failed or successful job.

**Stop**

stops an in-progress job.

**Note:** If you select **Stop**, your directive continues to display its In Progress status. In this situation, the directive is stopping, but it has not yet reached a suitable stopping point. Click **Refresh** periodically until the status changes to Stopped or reopen Run Status later to confirm the Stopped status.

## About Unsaved Jobs

If you run a directive without saving it, the directive is displayed in Run Status like any other directive. When processing stops on the unsaved directive, you can select **Open** from its Action menu. You can then edit and save the unsaved directive.

## About Incomplete Jobs

An incomplete job is one that you have stopped using the Action menu or one whose status is Failed. Depending on the type of the job and the point where execution ceased, log and code results might or might not be available.

---

## Saved Directives

### Introduction



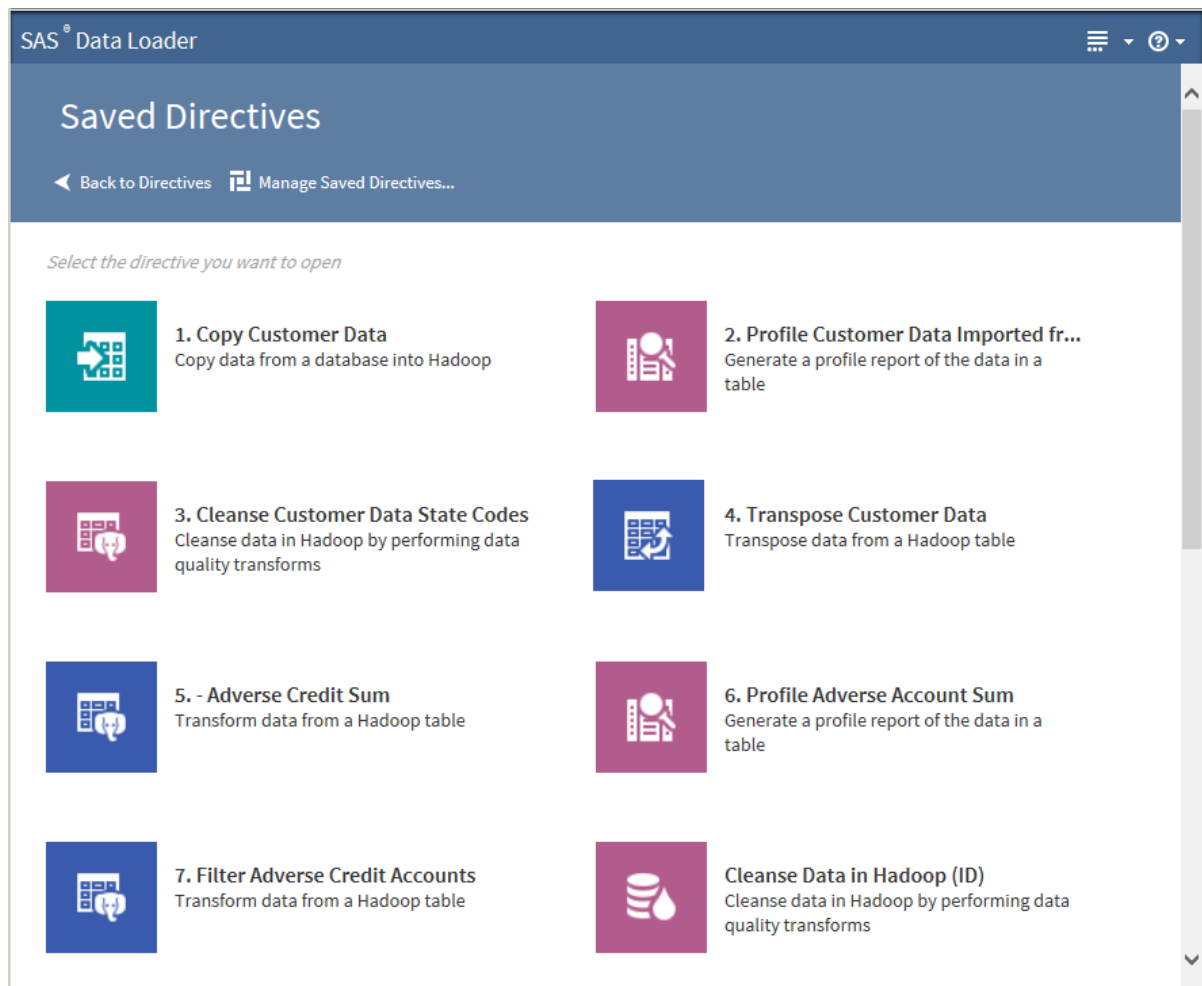
#### Saved Directives

Open a previously created directive to run, view or edit

Use Saved Directives to open, edit, and execute your saved directives. From the Saved Directives page, opening Manage Saved Directories enables you to open, duplicate, delete, refresh, or rename the selected directive.


### Opening Saved Directives

In the SAS Data Loader page, click the Saved Directives directive. A Saved Directives page similar to the following is displayed:



In the Saved Directives page, click a saved directive. The directive opens and can be edited or executed.

## Managing Saved Directories

In the Saved Directives page, click **Managed Saved Directives** . The Managed Saved Directives dialog box appears:

Name	Modified	Created
1. Copy Customer Data	Sep 25, 2014, 5:56:25 PM	Sep 25, 2014, 5:56:25 PM
2. Profile Customer Data	Sep 25, 2014, 6:28:04 PM	Sep 25, 2014, 6:28:04 PM
3. Cleanse Customer Data	Sep 25, 2014, 6:29:16 PM	Sep 25, 2014, 6:29:16 PM
4. Transpose Customer Data	Sep 26, 2014, 2:55:47 PM	Sep 26, 2014, 2:55:47 PM
5. - Adverse Credit Summary	Sep 29, 2014, 10:46:30 AM	Sep 29, 2014, 10:46:30 AM
6. Profile Adverse Account Summary	Sep 29, 2014, 11:14:00 AM	Sep 29, 2014, 11:14:00 AM
7. Filter Adverse Credit Accounts	Sep 29, 2014, 12:12:41 PM	Sep 29, 2014, 12:12:41 PM
Cleanse Data in Hadoop (ID)	Jan 27, 2015, 5:28:21 PM	Jan 27, 2015, 5:28:21 PM
Cleanse Data in Hadoop5	Nov 12, 2014, 3:51:10 PM	Nov 12, 2014, 3:51:10 PM
CQS Demo 1 - Copy Data to H...	Oct 22, 2014, 11:25:10 AM	Oct 22, 2014, 11:25:10 AM
CQS Demo 2 - Parse (Code Ac...	Oct 22, 2014, 12:00:29 PM	Oct 22, 2014, 12:00:29 PM
CQS Demo 3 - HiveQL (file type)	Oct 22, 2014, 1:32:26 PM	Oct 22, 2014, 1:32:26 PM

Copy data from a database into Hadoop

Close

Clicking the Action menu  enables the following actions:

**Open**

opens the selected directive.

**Duplicate**

duplicates the selected directive by opening a dialog box that enables you to assign a new name to the duplicated directive.

**Rename**

renames the selected directive.

**Delete**

deletes the selected directive.

**Refresh**

refreshes the selected directive, or, if no directive is selected, refreshes all of the saved directives in the list. Any duplicate, rename, or delete actions that you have taken are then reflected in the saved directives list.

## 7

## Client Administration

<i>Introduction</i> .....	107
<i>Update the vApp for SAS Data Loader</i> .....	107
<i>Troubleshoot the vApp Start Process</i> .....	108
<i>Update Kerberos Security on the vApp</i> .....	109
<i>Protect the vApp Directory</i> .....	109
<i>Troubleshoot Jobs</i> .....	109
Changing the Default Maximum Length for SAS Character Columns .....	109
Discover New Columns Added to a Source after Job Execution .....	110
Hive Limit of 127 Expressions per Table .....	110
Unsupported Hive Data Types and Values .....	110
<i>About Session Time-out</i> .....	111
<i>About Hadoop Client JAR Files and Client Configuration Files</i> .....	111
<i>Change the Version of Hadoop</i> .....	112
<i>Change the Hadoop Server Connection</i> .....	112
<i>Change the File Format of Hadoop Target Tables</i> .....	113
<i>Enable Logging inside the vApp</i> .....	115
<i>Manage Your License</i> .....	115
<i>Download Emergency SID Files</i> .....	116

### Introduction

The administrative tasks in this chapter apply after the initial installation and configuration of SAS Data Loader. To install and configure SAS Data Loader, refer to the *SAS Data Loader for Hadoop: Deployment Guide*.



### Update the vApp for SAS Data Loader

All of the client software for the SAS Data Loader for Hadoop runs inside the vApp. The vApp is a virtual machine that runs a separate operating system. All of the files that are accessed by the vApp are stored in a Shared Folder that

resides in the host operating environment. This architecture enables you to install vApp updates with one-button simplicity. Each update completely replaces the entire vApp. After the vApp update, there are no configuration or migration procedures. If you update the vApp, you might also see a link on the Information Center to notes that describe the changes in that release.

vApp updates require less than 15 minutes, given reasonable broadband capacity.

Follow these steps to check for the availability of vApp software updates, and to download and install updates.

- 1 Open the browser tab for the [SAS Data Loader: Information Center](#).
- 2 In the SAS Data Loader: Information Center, locate the **Notifications** section on the lower left.
- 3 To check to see whether a vApp update is available, click **Check for Updates**.
- 4 If a vApp update is available, open the Run Status directive to ensure that you have named and saved your jobs. If jobs are still running, click **Refresh**  to see their current status.
- 5 For any running directives, either wait for them to complete, or select the **Stop** option from the action menu .
- 6 Close the SAS Data Loader tab in the web browser.
- 7 Return to the SAS Data Loader: Information Center and click **Update**. The software update process stops the vApp, replaces the vApp, and then starts the new vApp in the VMware hypervisor.
- 8 When the SAS Data Loader: Information Center indicates that the vApp update is complete, click **Start SAS Data Loader**.

---

## Troubleshoot the vApp Start Process

If VMware Player Pro fails to start the vApp for SAS Data Loader, an error message states that Intel VT-x or AMD-v is not available. The message indicates that virtualization is either not supported or not configured in your BIOS (firmware.) Virtualization, also known as segmentation, enables the vApp to share display memory with the host operating system.

To resolve the error, determine your processor type, and then download and run a utility that enables virtualization, as described in the following steps:

- 1 Determine whether your computer has an Intel or AMD processor:
  - Press the Windows key and the R key on your keyboard at the same time. The Run dialog box appears.
  - In the **Open** field of the dialog box, type **msinfo32** and click **OK**.
  - In the System Information window, ensure that **System Summary** is selected in the left panel.



- In the right panel, find **System Type** and ensure that you have a 64-bit computer. Next, find **Processor** to determine the processor type (Intel or AMD.)
- 2 Download and use the tool that determines whether virtualization technology is supported on your processor:
    - Download the [Intel tool](#).
    - Download the [AMD tool](#).
  - 3 Visit [virtualization hardware extensions](#) page to enable Intel and AMD virtualization hardware extensions. To obtain information about how to navigate through your specific BIOS, contact the support site for the manufacturer of your computer.

For additional information about virtualization support, refer to the [VMware Knowledge Base](#).

---

## Update Kerberos Security on the vApp

If your Hadoop server is configured for Kerberos security, your vApp must also be configured. The required Kerberos configuration for the Hadoop cluster is described in the *SAS Data Loader for Hadoop: Administrator's Guide*. The required Kerberos configuration for the vApp is described in the *SAS Data Loader for Hadoop: Deployment Guide*.

---

## Protect the vApp Directory

SAS Data Loader contains encrypted passwords and other sensitive information. Do not share the vApp install directory with other users and protect it by making it accessible only to you.

---

## Troubleshoot Jobs

### Changing the Default Maximum Length for SAS Character Columns

Some directives use a SAS Server (SAS Workspace Server) to read or write tables. By default, the character columns for the input tables to such directives will be expanded to 32K in length. The affected directives are as follows:

- Transform Data in Hadoop
- Transpose Data in Hadoop
- Load Data to LASR
- Copy Data to Hadoop (when a data set from the SAS Server is selected as the input table)

- Copy Data from Hadoop (when the SAS Server is selected as the location for the target table)

For best performance, we recommend that you reduce the default length of the character columns for these input tables. Specify a value that is short enough to help performance but long enough to avoid truncating the data in character columns. For example, you could use a global option to reduce the column-length value for all new directive instances. You could use the advanced options for an individual directive to set a column-length value for that directive.

If you want to globally change the Maximum Length for SAS character columns for all new directive instances, go to the Directives page. Select **Configuration** from the Action menu. Select **General Preferences** from the Configuration window. Specify the desired length for SAS character columns.

If you want to change the Maximum Length for SAS character columns for a specific directive, go to the Directives page and open an appropriate directive. From the Source Table step, select **Advanced Options** from the Action menu at upper right. Specify the desired length for SAS character columns.

The only way to change the Maximum Length for SAS character columns for a specific directive is to set the appropriate advanced option for that directive. For example, you cannot change the global default for the Maximum Length for SAS character columns and have that dynamically applied to existing directives. The global option applies to new instances of directives only.

## Discover New Columns Added to a Source after Job Execution

When you add columns to a source table, any directives that need to use the new columns need to discover them. To make the new columns visible in a directive, open the Source Table page, click the source table again, and click **Next**. The new columns will then be available for use in the body of the directive, in a transformation or query, for example.

## Hive Limit of 127 Expressions per Table

Due to a limitation in the Hive database, tables can contain a maximum of 127 expressions. When the 128th expression is read, the directive fails and the SAS log receives a message similar to the following:

```
ERROR: java.sql.SQLException: Error while processing statement: FAILED:
Execution Error, return
      code 2 from org.apache.hadoop.hive.ql.exec.mr.MapRedTask
ERROR: Unable to execute Hadoop query.
ERROR: Execute error.
SQL_IP_TRACE: None of the SQL was directly passed to the DBMS.
```

The Hive limitation applies anytime a table is read as part of a directive. For SAS Data Loader, the error can occur in aggregations, profiles, when viewing results, and when viewing sample data.

## Unsupported Hive Data Types and Values

The Hive database in Hadoop identifies table columns by name and data type. To access a column of data, SAS Data Loader first converts the Hadoop column name and data type into its SAS equivalent. When the transformation is

complete, SAS Data Loader writes the data into the target table using the original Hadoop column name and data type.

If your target data is incorrectly formatted, then you might have encountered a data type conversion error.

The Hive database in Hadoop supports a Boolean data type. SAS does not support the Boolean data type in Hive at this time. Boolean columns in source tables will not be available for selection in SAS Data Loader.

The Bigint data type in Hive supports integer values larger than those that are currently supported in SAS. Bigint values that exceed +/-9,223,372,036,854,775,807 generate a stack overflow error in SAS.



---

## About Session Time-out

SAS Data Loader records periods of inactivity in the user interface. After a period of continuous inactivity, the current web page receives a session time-out warning message in a dialog box. If you do not provide input within three minutes after you receive the warning, the current web page is replaced by the Session Time-out page. You can restart your session by clicking the text **Return to the SAS Data Loader application**.

When a session terminates, any directives that you did not save or run are lost.

To open an unsaved directive that you ran before your session terminated, follow these steps:

- 1 Open the Run Status directive.
- 2 Locate the entry for your unsaved directive.
- 3 If the unsaved directive is still running, click the Refresh  button.
- 4 If the directive continues to run, either click **Stop** in the action menu , or wait for the completion of the run.
- 5 In the action menu, select **Open** to open the directive.
- 6 In the open directive, select **Save** from the title bar.

---

## About Hadoop Client JAR Files and Client Configuration Files

SAS Data Loader uses client Java Archive (JAR) files and XML client configuration files to connect the client to Hadoop. Both the JAR and XML files are produced and distributed by Hadoop vendors such as Cloudera and Hortonworks.

The client JAR files are obtained from vendors and pre-installed with SAS Data Loader. At install time, in the SAS Information Center, you select the version of Hadoop that is implemented on your Hadoop cluster. Your selection enables the appropriate JAR files. Within the available supported versions of Hadoop, you

can change versions at any time. To change versions, see the next topic “[Change the Version of Hadoop](#)”. New client JAR files for currently unsupported versions of Hadoop will be delivered in updates to SAS Data Loader.


The XML client configuration files are separate from the client JAR files. The XML files are copied from the Hadoop cluster onto the SAS Data Loader client during the installation of SAS Data Loader. The XML files are stored here:

`vApp-install-path\vApp-instance\SharedFolder\Configuration\HadoopConfig.`

---

## Change the Version of Hadoop

If you move to a different version of Cloudera or Hortonworks, and if that version is supported by SAS Data Loader, then follow these steps to select a different version.


- 1 Open the [SAS Data Loader: Information Center](#).
- 2 Click the Settings menu .
- 3 In the Settings window, change the value of the **Hadoop version** field.

---

## Change the Hadoop Server Connection

Follow these steps if you reconfigure your Hadoop server or move to a different Hadoop server.

**Note:** If your current Hadoop server uses Kerberos security, you cannot reconfigure your vApp to use a Hadoop server that does not use Kerberos security. In this case, you need to download a new vApp.

- 1 Open the [SAS Data Loader: Information Center](#).
- 2 Click the More menu  in the top right corner and select **Configuration**.
- 3 In the Configuration window, change as needed the following fields:
  - **Host** specifies the fully qualified network name of the Hadoop server.
  - **Port** specifies the port number that the server listens to for connections from SAS Data Loader.
  - **UserID** specifies the user name that SAS Data Loader uses to connect to the Hadoop server.
  - **Password** specifies the password that SAS Data Loader uses to connect to the Hadoop server.
  - **Oozie URL** specifies the URL to the Oozie Web Console, which is an interface to the Oozie server. The URL is similar to the following example: `http://host_name:port_number/oozie/`. Oozie is a workflow scheduler system that is used to manage Hadoop jobs.

- **Schema for temporary file storage** specifies an existing schema on the Hadoop server that stores temporary files and tables. To obtain the name of a non-default schema, open the data sources page in a directive such as Transform Data in Hadoop or Query a Table in Hadoop.

When your changes are complete, click **OK** to apply your changes and close the Configuration window.


---


## Change the File Format of Hadoop Target Tables

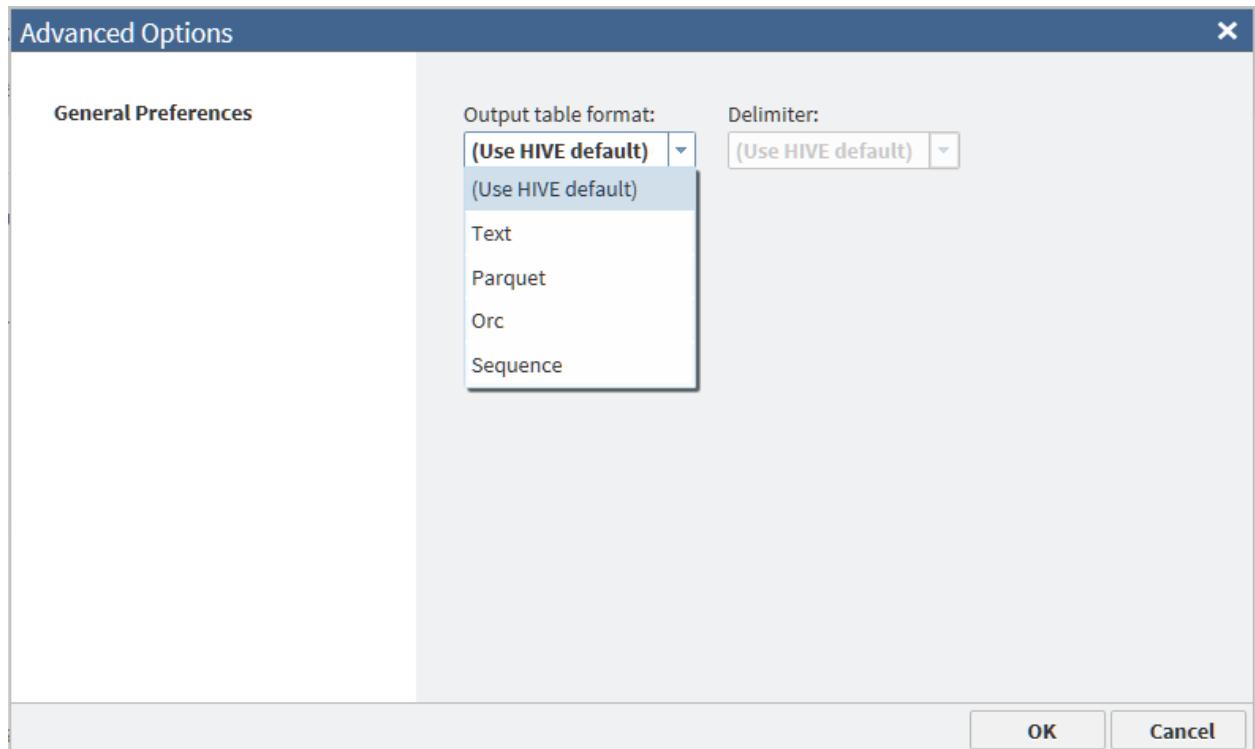
In Hadoop, tables are stored as one or more files in the Hadoop File System (HDFS). Each file is formatted according to the Output Table Format option, which is specified in each file. When you create a new target table in Hadoop, the Output Table Format option is set by the value of the **Output table format** field in SAS Data Loader.

You can change the default value of the **Output table format** field in the SAS Data Loader Configuration window. In any given directive, you can override the default value using the Action menu icon in the Target Table task.

The default format is applied to all new target tables that are created with SAS Data Loader. To override the default format in a new table or an existing table, you select a different format in the directive and run the job.

To change the default value of the **Output table format** field, click the More icon  in the top right corner of SAS Data Loader, and select **Configuration**. In the Configuration window, click **General Preferences** under **Hadoop Configuration**.

To override the default value of the **Output table format** field for a specific target table, open the directive, click the Action Menu icon  on the right side of the **Target Table** taskbar, and select **Advanced Options**.



The available values of the **Output table format** field are defined as follows:

**Use HIVE default**

specifies that the new target table receives the Output Table Format option value that is specified in HDFS. This is the default value for the **Output table format** field in SAS Data Loader.

**Text**

specifies that the new target table is formatted as a series of text fields that are separated by delimiters. For this option, you select a value for the **Delimiter** field. The default value of the **Delimiter** field is **(Use HIVE default)**. You can also select the value **Comma**, **Space**, **Tab**, or **Other**. If you select **Other**, then you enter a delimiter value. To see a list of valid delimiter values, click the question mark icon to the right of the **Delimiter** field.

**Parquet**

specifies the Parquet format, which is optimized for nested data. The Parquet algorithm is considered to be more efficient than using flattened nested name spaces.

**Orc**

specifies the Optimized Row Columnar format, which is a columnar format that efficiently manages large amounts of data in Hive and HDFS.

**Sequence**

specifies the SequenceFile output format, which enables Hive to efficiently run MapReduce. This format enables Hive to efficiently split large tables into separate threads.

Consult your Hadoop administrator for advice about output file formats. Testing might be required to establish the format that has the highest efficiency on your Hadoop cluster.

---

## Enable Logging inside the vApp


For debugging purposes, you can enable logging inside the vApp. To maintain performance, logging is not recommended under normal circumstances.

SAS recommends that you enable logging only when you are directed to do so by your SAS Technical Support representative.

Inside the vApp, log files are generated by a SAS Object Spawner and a series of SAS Workspace Servers. The SAS Object Spawner creates a new instance of the SAS Workspace Server for each HTTP session. When logging is enabled, the SAS Object Spawner generates the log files `ObjectSpawner_console_vsasmaster.log` and `ObjectSpawner_YYYY-MM-DD_localhost_PID.log`. The SAS Workspace generates the log file `SASApp_WorkspaceServer_YYYY-MM-DD_localhost_PID.log`.

The log files are stored in the following location: `vApp-path\vApp-instance\Shared Folder\Logs`.

Follow these steps to enable or disable logging inside the vApp:

- 1 Check the [Run Status](#) directive to ensure that no directives are running. This is important because the SAS Object Spawner and other services restart when you enable logging. The same services also restart when you disable logging.
- 2 Open the [SAS Data Loader: Information Center](#).
- 3 Click the Settings icon .
- 4 In the Settings window, click **Advanced**.
- 5 To activate logging, click **Turn logging on (for debugging only)**. To deactivate logging, click **Turn logging off**. Click **OK**.
- 6 In the Applying Settings Changes window, click **Yes** to confirm your selection. Click **No** to make no change.

---

## Manage Your License


During installation, your installer identified the local storage location of a SAS installation data file (SID). The SID file contains your license. The SID file is delivered as an attachment to the Software Order Email.

If your initial license has yet to be applied to SAS Data Loader, follow the renewal steps below.

To check to see whether a license has been identified, open the SAS Information Center and click the Settings icon in the top right corner. In the Settings window, if the check box Apply New License has been selected, then your license has been identified to SAS Data Loader.

The license remains valid for a year after the receipt of the Software Order Email.

Ten days before the expiration of your license, the Configuration menu in SAS

Data Loader  displays a message that states the number of days available before the expiration of the license. On the expiration date, the message states that your license has expired.


Beginning on the day your license expires, the SAS Information Center web page displays the following message each time you open SAS Data Loader:

```
SETINIT Expiration
Your SETINIT for this product is nearing expiration.
```

The word SETINIT is a name that refers to the SAS license.

**Note:** The expiration message is not displayed in the SAS Information Center when you use the Firefox web browser.

When you begin to receive expiration messages, you should contact your SAS Installation Representative to renew your license. Upon renewal, you will receive a Renewal Order Email that contains a new SID file. Follow these steps to identify the new SID file in SAS Data Loader:

- 1 Save the SID file from your Renewal Order Email to a directory on the computer that hosts the vApp.
- 2 If necessary, open the [SAS Data Loader: Information Center](#).
- 3 If SAS Data Loader is currently open in a web browser, first close the web browser tab to stop SAS Data Loader. (Doing so will not have a negative impact on any running directives.) Second, open the SAS Information Center by clicking the icon on your desktop. Another way to open the SAS Information Center is to enter into a web browser the web address that is displayed in the VMware Player Plus window.
- 4 In the SAS Information Center, click the Settings icon  in the top right corner.
- 5 In the Settings window, click **Apply New License**, and then click **Browse**.
- 6 In your file browser, navigate to the directory that contains the license file, click the license file, and click **Open**.
- 7 In the Settings window, click **Yes**.
- 8 To begin using the new license, simply open SAS Data Loader for Hadoop in a web browser.

You have now renewed your license for SAS Data Loader for Hadoop. The new license will remain valid for the time period that is specified in your Renewal Order Email.

---

## Download Emergency SID Files

Follow these steps to download a temporary SID file that will extend the use of your licensed SAS software products for six days:



- 1 In a web browser, open the SAS Install Center, at <http://support.sas.com/documentation/installcenter/index.html>.
- 2 Under **Site and Account Data** on the right side of the page, select **Request a Temporary License Extension**. You can also select **Resend the SAS Installation Data**.
- 3 After you receive your temporary SID file, identify that file to SAS Data Loader as described in ["Manage Your License"](#).



## Recommended Reading

- *SAS Data Loader for Hadoop: vApp Deployment Guide*
- *SAS Data Loader for Hadoop: Administrator's Guide*
- *SAS 9.4 DS2 Language Reference*
- *SAS/ACCESS for Relational Databases: Reference*
- *SAS 9.4 In-Database Products: Administrator's Guide*
- *SAS Quality Knowledge Base for Contact Information 23: Installation and Configuration* (see the online Help for usage information)

For a complete list of SAS publications, go to [sas.com/store/books](http://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  
E-mail: [sasbook@sas.com](mailto:sasbook@sas.com)  
Web address: [sas.com/store/books](http://sas.com/store/books)



# Index

## A

administration [107](#)

## C

clear Run Status entries [103](#)  
 Cloudera [112](#)  
 columns, discover new [110](#)  
 Configuration window [96](#), [112](#)

## D

directives  
   incomplete [104](#)  
   troubleshoot [110](#)  
   unsaved [104](#)  
 Directives page [14](#)

## H

Hadoop  
   client JAR files [111](#)  
   client XML configuration files [111](#)  
   server connection [112](#)  
 Hive  
   data types [110](#)  
   limit on expressions [110](#)  
   maximum integer value [111](#)  
 Hortonworks [112](#)

## I

incomplete directives [104](#)

## L

LASR Server Configuration window  
[97](#)

license management [115](#)  
 Load Data to LASR directive [95](#)  
 logging [115](#)

## M

More menu [112](#)

## P

prerequisites  
   SAS LASR Analytic Server [95](#)  
 Profile Data directive [55](#)  
 Profile, Saved Reports [62](#)

## Q

Query or Join Data in Hadoop  
 directive [32](#), [42](#)

## R

Run Status directive [102](#)

## S

SAS Data Loader  
   architecture [2](#)  
 SAS Data Quality Accelerator for  
   Hadoop [2](#)  
 SAS Information Center [112](#)  
   close [9](#)  
   open [9](#)  
 SAS LASR Analytic Server [95](#)  
   configure SSH keys [98](#)  
 SAS LASR Server  
   LASR Server Configuration window  
     [97](#)  
 SAS Table Viewer [15](#)  
 SAS Visual Analytics [95](#)

SAS/ACCESS for Hadoop [2](#)  
Saved Directives [104](#)  
Saved Profile Reports directive [62](#)  
session time-out [111](#)  
Settings menu [112](#)  
shared folder [2](#), [112](#)  
Summarize Rows transformation [50](#)

## T

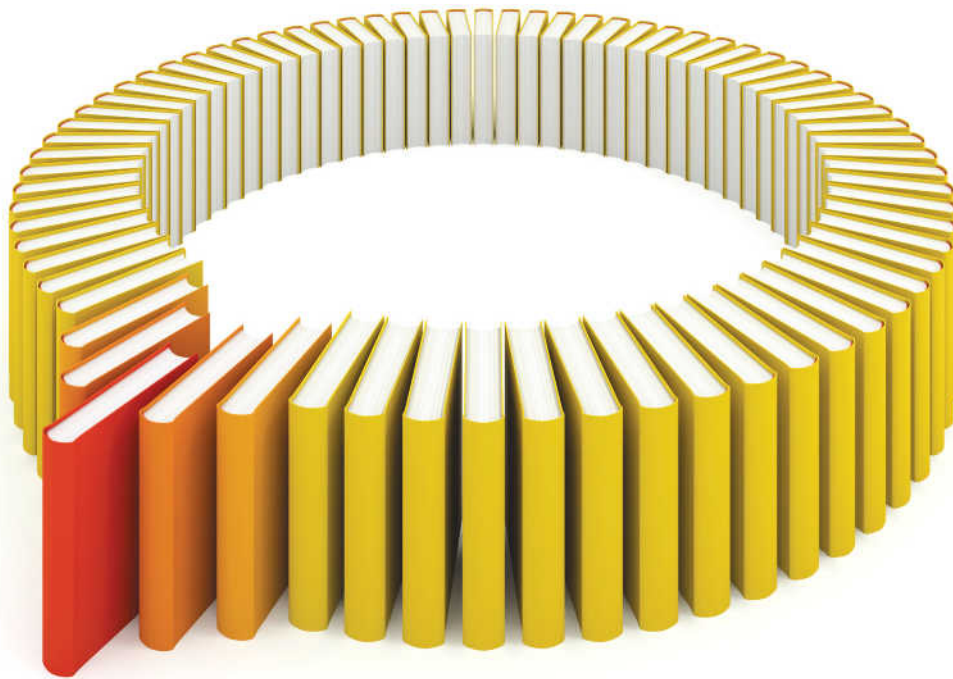
Transpose Data in Hadoop directive  
[51](#)  
troubleshoot directives [110](#)

## U

unsaved directives [104](#)  
update SAS Data Loader [107](#)

## V

vApp  
    how it works [2](#)  
    security [109](#)  
VMware Player Plus [107](#)



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

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



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

